Econometría Básica

Aplicaciones con EVIEWS, STATA, SAS y SPSS

El objetivo de este libro es la presentación de las técnicas econométricas básicas y su tratamiento con las herramientas más adecuadas de cálculo automatizado. Se utilizan los paquetes de software EVIEWS, STATA, SAS y SPSS para abordar de modo sencillo el trabajo econométrico.

El primer bloque de contenido se ocupa del modelo lineal de regresión múltiple y de toda su problemática, incluyendo herramientas para la detección y tratamiento de la autocorrelación, heteroscedasticidad, multicolinealidad, normalidad residual, linealidad, observaciones influyentes, errores de especificación, exogeneidad y regresores estocásticos.

Un segundo bloque trata los modelos de elección discreta, recuento, censurados, truncados y de selección muestral, haciendo hincapié en los modelos Logit, Probit, Tobit, Poisson, binomial negativa y corrección del sesgo de selección mediante la estimación de Heckman.

El tercer bloque aborda el análisis univariante de series temporales a través de la metodología Box Jenkins para modelos ARIMA, el tratamiento de los modelos de intervención y los modelos univariantes de la función de transferencia.

El último bloque se ocupa de los modelos del análisis de la varianza y la covarianza, el modelo lineal general y los modelos mixtos.

Los capítulos se inician con la exposición de los conceptos y notas teóricas adecuadas, para resolver a continuación una variedad de ejercicios que cubran los conceptos expuestos. Se trata de recopilar la mayor parte de los conceptos econométricos e ilustrarlos con la práctica a través de las herramientas de software adecuadas.

César Pérez López es licenciado en Ciencias Matemáticas y en Ciencias Económicas y trabaja actualmente como Vocal Asesor en el Instituto de Estudios Fiscales (IEF). Pertenece al Cuerpo Superior de Estadísticos del Estado y al Cuerpo Superior de Sistemas y Tecnologías de la Información de la Administración del Estado. Asimismo, es profesor asociado en el Departamento de Estadística e Investigación Operativa III de la Universidad Complutense de Madrid.

Garceta grupo editorial

www.garceta.es

ISBN 978-84-1545-202-7

9 788415 452027

Econometría Básica

Aplicaciones con EVIEWS, STATA, SAS y SPSS

César Pérez

330.015195
P438e

Garceta grupo editorial

Econometría Básica

Aplicaciones con EVIEWS, STATA, SAS y SPSS

César Pérez

BIBLIOTECA UNIVERSIDAD JAVERIANA-SEC CALI
4 0000 00210617 9

César Pérez López
Instituto de Estudios Fiscales (IEF)
Universidad Complutense de Madrid

Econometría Básica

Aplicaciones con EVIEWS, STATA, SAS y SPSS

Garceta
grupo editorial

Econometría Básica. Aplicaciones con EVIEWS, STATA, SAS y SPSS

César Pérez López

ISBN: 978-84-1545-202-7
IBERGARCETA PUBLICACIONES, S.L., Madrid 2012

Edición: 1.ª
Impresión: 1.ª
N.º de páginas: 630
Formato: 17 x 24 cm

Materia CDU: 33 Economía en general

Reservados los derechos para todos los países de lengua española. De conformidad con lo dispuesto en el artículo 270 y siguientes del código penal vigente, podrán ser castigados con penas de multa y privación de libertad quienes reprodujeren o plagiaren, en todo o en parte, una obra literaria, artística o científica fijada en cualquier tipo de soporte sin la preceptiva autorización. Ninguna parte de esta publicación, incluido el diseño de la cubierta, puede ser reproducida, almacenada o trasmitida de ninguna forma, ni por ningún medio, sea éste electrónico, químico, mecánico, electro-óptico, grabación, fotocopia o cualquier otro, sin la previa autorización escrita por parte de la editorial.

Diríjase a CEDRO (Centro Español de Derechos Reprográficos), www.cedro.org, si necesita fotocopiar o escanear algún fragmento de esta obra.

COPYRIGHT © 2012 IBERGARCETA PUBLICACIONES, S.L.
info@garceta.es

Econometría Básica. Aplicaciones con EVIEWS, STATA, SAS y SPSS

César Pérez López

1.ª edición, 1.ª impresión
OI: 275/2015
ISBN: 978-84-1545-202-7
Deposito Legal: M-7214-2012
Imagen de cubierta: © Jose Ignacio Soto_Fotolia.com

Impresión: Print House, marca registrada de Coplar, S.A.

IMPRESO EN ESPAÑA - PRINTED IN SPAIN

Nota sobre enlaces a páginas web ajenas: Este libro puede incluir referencias a sitios web gestionados por terceros y ajenos a IBERGARCETA PUBLICACIONES, S.L., que se incluyen sólo con finalidad informativa. IBERGARCETA PUBLICACIONES, S.L., no asume ningún tipo de responsabilidad por los daños y perjuicios derivados del uso de los datos personales que pueda hacer un tercero encargado del mantenimiento de las páginas web ajenas a IBERGARCETA PUBLICACIONES, S.L., y del funcionamiento, accesibilidad y mantenimiento de los sitios web no gestionados por IBERGARCETA PUBLICACIONES, S.L., directamente. Las referencias se proporcionan en el estado en que se encuentran en el momento de publicación sin garantías, expresas o implícitas, sobre la información que se proporcione en ellas.

330.015195
P438e

CONTENIDO

Introducción ..... *XV*

Capítulo 1. Modelo lineal de regresión múltiple. Hipótesis, Estimación, inferencia y predicción *1*

Modelo lineal de regresión múltiple 1
Hipótesis en el modelo lineal 2
Hipótesis relativas a la perturbación aleatoria 2
Hipótesis relativas a los regresores 3
Hipótesis relativas al vector de parámetros 3
Hipótesis relativas a la forma funcional 3
Estimación del modelo lineal por mínimos cuadrados ordinarios MCO 3
Estimación del modelo lineal por máxima verosimilitud 6
Inferencia en el modelo por mínimos cuadrados ordinarios MCO 7
Predicciones 13
Predicciones puntuales y en media 14
Capacidad predictiva de un modelo 14
Selección de modelos de regresión 16
Análisis de los residuos 17
Modelo lineal con restricciones 18
Regresión con variables cualitativas: variables ficticias 21
Modelos de regresión con variables cualitativas 21
Variables ficticias en el análisis estacional 23
Variables ficticias en la regresión por tramos 24

Capítulo 2. Modelo lineal de regresión múltiple. Herramientas de software *45*

EVIEWS y el trabajo básico con le modelo de regresión múltiple 45
SPSS y el trabajo básico con el modelo de regresión múltiple 50
SAS y el trabajo básico con el modelo de regresión múltiple 53
STATA y el trabajo básico con el modelo de regresión múltiple 74
STATA y el trabajo básico con el modelo de regresión múltiple a través de menús 76

Capítulo 3. Autocorrelación, heteroscedasticidad, multicolinealidad, no linealidad y normalidad *115*

Modelos con autocorrelación 115
Detección de la autocorrelación 116

Estadístico DW de Durbin-Watson 117
Estadístico h de Durbin 117
Estadístico D4 de Wallis (datos trimestrales) 118
Razón de Von Neumann 118
Prueba de Breusch-Godfrey 119
Soluciones para la autocorrelación 119
Método de mínimos cuadrados generalizados 120
Métoso de Prais-Winsten 120
Método iterativo de Cochrane.Orcutt 120
Método de Durbin 121
Modelos con heteroscedasticidad 122
Detección de la heteroscedasticidad 122
Contraste de Goldfeld-Quandt 122
Contraste de Glesjer 123
Contraste W de White 123
Contraste RESET de Ramsey 123
Contraste de Breush-Pagan 123
Heteroscedasticidad condicional. Contrastes ARCH y GARCH 124
Contraste de rangos 124
Soluciones para la heteroscedasticidad 125
Mínimos Cuadrados Generalizados y Mínimos Cuadrados Ponderados ... 125
Ajuste de White 126
Soluciones para la heteroscedasticidad condicional: modelos ARCH y GARCH 126
Multicolinealidad 128
Detección de la multicolinealidad 128
Soluciones para la multicolinealidad 128
Normalidad residual 129
El problema de la falta de normalidad en los residuos 129
Soluciones para la falta de normalidad en los residuos 131
No linealidad y errores de especificación 131
Error de especificación en la selección de las variables explicativas 133
Error de especificación en la forma funcional 134
Exogeneidad y regresores estocásticos 135
El método de las variables instrumentales 136
El estimador de mínimos cuadrado en dos etapas MC2E 137
El contraste de Hausman 137
Análisis de la influencia 138
Influencia a priori (Leverage) 138
Distancia de Mahalanobis 140
Criterio de los residuos eliminados 140
Influencia a posteriori: estadísticos de Cook y DFFITS 141

Capítulo 4. Herramientas para tratar autocorrelación, Heteroscedasticidad y otros problemas 165

Tratamiento de la autocorrelación y heteroscedasticidad con Eviews 165
Eviews y los modelos ARCH Y GARCH 171
Endogeneidad, variables instrumentales y mínimos cuadrados en dos etapas con Eviews 177
Errores de especificación con Eviews. Variables omitidas y redundantes .. 182
Errores de especificación en la forma funcional con Eviews 186
SPSS y modelos con regresores estocásticos. Variables instrumentales y M.C. en dos fases 190
SPSS y modelos con heteroscedasticidad y multicolinealidad. Mínimos cuadrados ponderados 195
SAS y la multicolinealidad, autocorrelación, heteroscedasticidad, valores influyentes y errores de especificación 199
SAS y los modelos ARCH Y GARCH 206
STATA y la multicolinealidad, autocorrelación, heteroscedasticidad, errores de especificación y observaciones influyentes 210
STATA y la multicolinealidad, autocorrelación, heteroscedasticidad, errores de especificación y observaciones influyentes a través de menús ... 216

Capítulo 5. Modelos Logit, Probit, Tobit, truncados, recuento, censurados y de selección muestral. Herramientas 253

Modelos variable dependiente limitada 253
Modelos de elección discreta 254
Modelos de elección discreta binaria 255
Modelo lineal de probabilidad 255
Modelos Probit y Logit 256
Modelos de elección múltiple 260
Modelo Logit Multinomial 261
Modelo Logit Condicional 262
Modelo Logit Anidado 264
Modelo Probit Multinomial 264
Modelos Logit y Probit otdenados 265
Modelo de datos de recuento 266
Modelo de Regresión de Poisson 267
Modelo de Regresión de Binomial Negativa 269
Modelo de Regresión Exponencial 270
Modelo de Regresión Normal 270
Modelos censurados. El modelo Tobit 270
Interpretación de los coeficientes en el modelo Tobit 271

Efectos parciales en el modelo Tobit 272
Modelo Tobit con censura en los datos 273
Estimación máximo verosímil (MV) del modelo Tobit censurado 274
Estimación por mínimos cuadrados ordinarios en dos etapas del modelo Tobit censurado 274
Selección muestral. Modelos truncados 275
Estimación del modelo Tobit truncado 276
Efectos parciales en el modelo Tobit truncado 276
Estimación máximo verosímil (MV) del modelo Tobit truncado 277
Estimación por el método de Amemiya en dos etapas 277
Corrección de la selección muestral. Estimación bietápica de Heckman o Heckit 278
SPSS y la regresión logística binaria 279
SPSS y el modelo Probit 287
SPSS y el modelo Logit multinomial 290
SAS y la regresión logística. Proc LOGISTIC 296
SAS y el modelo Probit. Procedimiento PROBIT 303
SAS y el modelo Tobit de regresión censurada. Procedimiento LIFEREG 307
Modelos de variable dependiente limitada con Eviews: MLP, Logit y Probit 310
Modelos de recuento con Eviews: Poisson, binomial negativa y exponencial 318
Modelos Tobit censurado y truncado con Eviews. Método de Heckman y Ratio de Mills 322
Modelos de variable dependiente limitada con STATA: Logit y Probit 330
Modelos Tobit censurado y truncado con STATA. Método de Heckman y Ratio de Mills 333
Modelo de Poisson con STATA 335

Capítulo 6. Análisis univariante de series temporales. Modelos ARIMA, intervención y función de transferencia 351

Series temporales 351
Descomposición clásica de una serie temporal 352
Tendencia de una serie temporal: ajuste analítico, medias móviles y diferencias 353
Variaciones estacionales: medias móviles, diferencias, diferencias estacionales y variables ficticias 355
Variaciones cíclicas 358
Predicción y suavizado de series temporales, métodos autoproyectivos deterministas 360
Suavizado por medias móviles 360
Suavizado lineal de Holt 361
Suavizado exponencial de Brown 361

Suavizado estacional de Winters 362
Predicciones incondicionales estocásticas 363
Modelos ARIMA: Primeros conceptos 363
Series temporales y procesos estocásticos. Características 364
Procesos estocásticos estacionarios. Funciones de autocorrelación y autocorrelación parcial 365
Series temporales estacionarias. Detección de la estacionariedad 366
Modelos autorregresivos AR(p) 369
Modelos de medias móviles MA(q) 371
Modelos ARMA(p,q) 375
Modelos ARIMA(p,d,q) 377
La metodología de Box Jenkins en modelos ARIMA 379
Identificación de modelos ARIMA 380
Estimación de modelos ARIMA(p,d,q) 382
Diagnóstico, validación o contraste de modelos ARIMA(p,d,q) 382
Predicción en Modelos ARIMA 384
Series temporales estacionales. Detección de la estacionalidad 385
Modelos estacionales puros 387
Modelos autorregresivos estacionales AR(P)s 387
Modelos de medias móviles estacionales MA(Q)s 388
Modelos estacionales ARMA(P,Q)s 390
Modelos ARIMA(P,D,Q)s estacionales puros 391
Identificación de modelos estacionales puros 392
Modelos estacionales generales 394
Modelos estacionales generales con parte regular autorregresiva. Identificación 394
Modelos estacionales generales con parte regular de media móvil. Identificación 397
Identificación de modelos estacionales ARIMA(p,d,q)(P,D,Q)s 401
Estimación de modelos estacionales ARIMA(p,d,q)(P,D,Q)s 403
Validación de modelos estacionales ARIMA(p,d,q)(P,D,Q)s y predicción 404
Modelos de intervención 405
Variables escalón e impulso 405
Modelo de intervención general 408
Identificación de modelos de intervención 410
Valores atípicos (Outliers) 411
Tipos de outliers 412
Outliers aditivos (AO) 412
Outliers innovacionales (IO) 413
Outliers de cambio de nivel (LS) 413
Outliers de cambio temporal (TC) 414
Modelo univariante de la función de transferencia 415

Identificación, estimación y validación del modelo de la función de transferencia 417
Etapas de la identificación, estimación y validación del modelo de la función de transferencia 421
Modelos de la función de transferencia estacionales 422

Capítulo 7. Herramientas para el análisis univariante de series temporales.. *441*

Eviews y la identificación, estimación, validación y predicción de modelos ARIMA(p,d,q)(P,D,Q)s 441
Eviews y los modelos ARIMA y de intervención 449
Tramo/Seats 449
Eviews y los métodos autoproyectivos deterministas: alisados exponenciales y de Holt_Winters 456
SPSS y la identificación, estimación, diagnosis y predicción de modelos ARIMA(p,d,q)(P,D,Q)s 457
Predicciones incondicionales deterministas con el modelizador de SPSS. Suavizado 469
SPSS y los modelos ARIMA con intervención 478
SAS y la identificación, estimación, validación y predicción de modelos ARIMA(p,d,q)(P,D,Q)s 485
Procedimiento ARIMA 485
SAS y los modelos ARIMA de intervención y de función de transferencia 495
STATA y los modelos ARIMA(p,d,q)(P,D,Q)s 516
Suavizado con STATA 520

Capítulo 8. Modelos del análisis de la varianza y la covarianza. Modelo Lineal General y modelos mixtos *541*

Modelos del análisis de la varianza y la covarianza 541
Modelos ANOVA de la varianza simple 543
Modelo ANOVA con un solo factor. Efectos fijos y aleatorios 543
Modelo ANOVA con dos factores. Efectos fijos, aleatorios y mixtos 545
Modelo ANOVA con tres factores 547
Modelos ANCOVA de la covarianza simple 547
Modelo con un factor y un covariante 548
Modelo con dos factores y un covariante 548
Modelos con dos factores y dos covariantes 549
Análisis multivariante de la varianza (MANOVA) 549
Análisis multivariante de la varianza con un factor 549
Análisis multivariante de la varianza con dos factores 550
Análisis multivariante de la covarianza (MANCOVA) 551

Modelo Lineal General (GLM) 551
Modelos lineales mixtos 551
SPSS y los modelos ANOVA y ANCOVA univariantes de uno y varios factores 553
SPSS y la estimación de las componentes de la varianza en modelos ANCOVA de efectos mixtos 563
SPSS y los modelos MANOVA y MANCOVA multivariantes de uno y varios factores 565
SPSS y los modelos lineales mixtos 573
Análisis de la varianza y la covarianza con SAS. Procedimiento GLM 576
Componentes de la varianza en SAS. Procedimiento VARCOMP 584
SAS y los modelos mixtos. PROC MIXED 586
STATA y el análisis de la varianza-covarianza. El modelo GLM y los modelos mixtos 592

A mis apoyos cercanos

INTRODUCCIÓN

El objetivo de este libro es la presentación de las técnicas econométricas básicas, tanto clásicas como modernas, y su tratamiento con las herramientas más adecuadas de cálculo automatizado. Se utilizarán los paquetes de software más habituales, como son EVIEWS, STATA, SAS y SPSS, para abordar de modo sencillo el trabajo econométrico. Los capítulos se inician con la exposición de los conceptos y notas teóricas adecuadas, para resolver a continuación una variedad de ejercicios que cubran los conceptos expuestos. No se trata, por tanto, de hacer una exposición teórica completa con demostraciones, sino más bien de recopilar la mayor parte de los conceptos econométricos e ilustrarlos con la práctica a través de las herramientas de software adecuadas.

Partiremos de la consideración de las técnicas de modelización econométrica como un subconjunto de las técnicas más generales de análisis de datos, que engloban tanto *técnicas predictivas* enfocadas a la *modelización y clasificación ad hoc*, como *técnicas descriptivas* enfocadas generalmente a la *clasificación post hoc* y otro tipo de técnicas variadas.

Las *técnicas predictivas*, base de la econometría, especifican el modelo para los datos de acuerdo a un conocimiento teórico previo recogido en la teoría económica. Una vez identificado el modelo teórico para los datos, se procede a su estimación debiendo ser posteriormente contrastado antes de aceptarlo como válido. Posteriormente ya puede utilizarse el modelo para predecir. Tenemos así las cuatro fases típicas de la modelización econométrica: identificación, estimación, diagnosis y predicción. Podemos incluir entre las técnicas predictivas todos los tipos de regresión, series temporales, análisis de la varianza y covarianza, modelos de diseño de experimentos, análisis discriminante, árboles de decisión y redes neuronales.

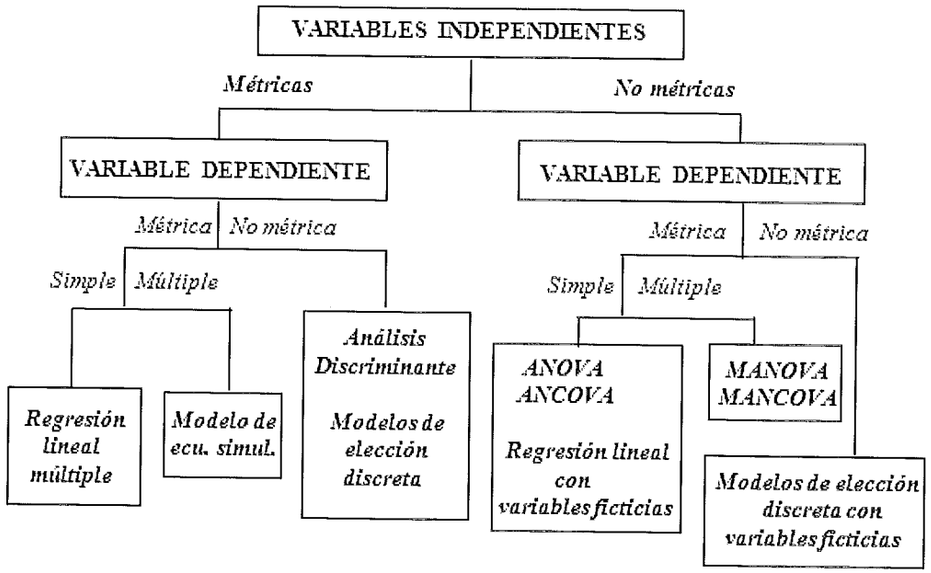
Pero, tanto los árboles de decisión, como las redes neuronales y el análisis discriminante son a su vez *técnicas de clasificación* que pueden extraer perfiles de comportamiento o clases, siendo el objetivo construir un modelo que permita clasificar cualquier nuevo dato. Los árboles de decisión permiten clasificar los datos en grupos basados en los valores de las variables. El mecanismo de base consiste en elegir un atributo como raíz y desarrollar el árbol según las variables más significativas. De esta forma se puede realizar en cierto modo perfilado y segmentación de datos.

Alternativamente, en las *técnicas descriptivas* no se asigna ningún papel predeterminado a las variables. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean automáticamente partiendo del reconocimiento de patrones. En este grupo se incluyen las técnicas de *clustering* y segmentación (que también son técnicas de clasificación en cierto modo), las técnicas de asociación y dependencia, las técnicas de análisis exploratorio de datos y las técnicas de reducción de la dimensión (factorial, componentes principales, correspondencias, etc.) ya vistas en la fase de transformación. El esquema siguiente clasifica las técnicas de análisis de datos.

- *Técnicas de análisis de datos*
  - *Predictivas*
    - *Modelización*
      - *Modelos de regresión*
      - *Modelos de elección discreta*
      - *Análisis de la varianza - covarianza*
      - *Modelos de diseño de experimentos*
      - *Análisis de series temporales*
      - *Modelos de ecuaciones simultáneas*
      - *Redes neuronales*
    - *Clasificación ad hoc*
      - *Discriminante*
      - *Árboles de decisión*
  - *Descriptivas*
    - *Clasificación post hoc*
      - *Clustering*
      - *Segmentación*
    - *Asociación*
    - *Dependencia*
    - *Reducción de la dimensión*
    - *Análisis exploratorio*

Se observa que las *técnicas de clasificación* pueden pertenecer tanto al grupo de técnicas predictivas (discriminante, árboles de decisión y redes neuronales) como a las descriptivas (clustering y segmentación). Las técnicas de clasificación predictivas suelen denominarse *técnicas de clasificación ad hoc* ya que clasifican individuos u observaciones dentro de grupos previamente definidos. Las técnicas de clasificción descriptivas se denominan *técnicas de clasificación post hoc* porque realizan clasificación sin especificación previa de los grupos. Por otro lado, las redes neuronales pueden utilizarse tanto para la modelización como para la clasificación.

Hemos visto que la clasificación de las técnicas de análisis de datos discrimina entre la existencia o no de variables explicativas y explicadas. Si existe una dependencia entre las variables explicadas y sus correspondientes variables explicativas, que pueda plasmarse en un ***modelo***, estamos ante las *técnicas predictivas* o *métodos explicativos*. Este tipo de técnicas de análisis de la dependencia pueden clasificarse en función de la naturaleza métrica o no métrica de las variables independientes y dependientes como sigue:



El ***análisis de la regresión múltiple*** es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas) también métricas. El objetivo esencial del análisis de la regresión múltiple es utilizar las variables independientes, cuyos valores son conocidos, para predecir la única variable criterio (dependiente) seleccionada por el investigador.

La expresión funcional del análisis de la regresión múltiple es la siguiente:

$$y = F(x_1, x_2, \cdots, x_n)$$

donde inicialmente, tanto la variable dependiente y como las independientes x_i son métricas. Asimismo la regresión múltiple admite la posibilidad de trabajar con variables independientes no métricas si se emplean variables ficticias (***modelos de regresión con variables ficticias***) para su transformación en métricas.

Los ***modelos de ecuaciones simultáneas*** constituyen una técnica estadística utilizada para analizar la relación entre múltiples variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) también métricas. El objetivo esencial es utilizar las variables independientes, cuyos valores son conocidos, para predecir las variables criterio (dependientes) seleccionadas por el investigador.

La expresión funcional es la siguiente:

$$\vec{G}(y_1, y_2, \cdots, y_n) = \vec{F}(x_1, x_2, \cdots, x_n)$$

donde inicialmente, tanto las variables dependientes y_i como las independientes x_i son métricas. Se observa que este modelo es una ampliación del modelo de regresión múltiple al caso de varias variables dependientes.

El ***análisis discriminante*** es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) no métrica (categórica) y varias variables independientes (o exógenas) métricas. El objetivo esencial del análisis discriminante es utilizar los valores conocidos de las variables independientes para predecir con qué categoría de la variable dependiente se corresponden. Así podremos predecir en qué categoría de riesgo crediticio se encuentra una persona, el éxito de un producto en el mercado, etc.

La expresión funcional del análisis discriminante es la siguiente:

$$y = F(x_1, x_2, \cdots, x_n)$$

donde la variable dependiente y es no métrica y las variables independientes son métricas. Se trata por tanto de un caso particular del análisis de regresión múltiple. Formalmente podríamos decir que, el análisis discriminante es una técnica de clasificación que permite agrupar a los elementos de una muestra en dos o más categorías diferentes, predefinidas en una variable dependiente no métrica, en función de una serie de variables independientes métricas combinadas linealmente.

En el análisis discriminante, para valores dados de las variables independientes hemos de predecir la probabilidad de pertenencia a una categoría o clase de la variable dependiente (por ejemplo, probabilidad de que un individuo compre un producto o devuelva un crédito según algunas variables medidas en él). Los ***modelos de elección discreta*** tienen la misma naturaleza que el modelo discriminante, pero ahora lo que se predice es la probabilidad de pertenencia a una categoría (clase) para valores dados de las variables dependientes. Por tanto, los modelos de elección discreta predicen directamente la probabilidad de ocurrencia de un suceso que viene definido por los valores de las variables independientes. Como los valores de una probabilidad están entre cero y uno, las predicciones realizadas con los modelos de elección discreta deben estar acotadas para que caigan en el rango entre cero y uno. El modelo general que cumple esta condición es un caso particular del modelo de regresión múltiple que se denomina ***modelo lineal de probabilidad***, y tiene la forma funcional:

$$P_i = F(x_i, \beta) + u_i$$

Se observa que si F es la función de distribución de una variable aleatoria, entonces P varía entre cero y uno.

En el caso particular en que la función F es la función logística estaremos ante el ***modelo Logit o Regresión Logística***, cuya forma funcional será la siguiente:

$$P_i = F(x_i, \beta) + u_i = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} + u_i$$

En el caso particular en que la función F es la función de distribución de una normal unitaria estaremos ante el ***modelo Probit***, cuya forma funcional será la siguiente:

$$P_i = F(x_i, \beta) + u_i = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x_i\beta} e^{-\frac{t^2}{2}} dt + u_i$$

El ***análisis de la varianza simple*** es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas) no métricas. El objetivo esencial de los modelos del análisis de la varianza es determinar si diversas muestras proceden de poblaciones con igual media. Los valores no métricos de las variables independientes determinarán una serie de grupos en la variable dependiente. De modo que el modelo ANOVA mide la significación estadística de las diferencias entre las medias de los grupos determinados en la variable dependiente por los valores de las variables independientes.

La expresión funcional del modelo del análisis de la varianza simple ANOVA es la siguiente:

$$y = F(x_1, x_2, \cdots, x_n)$$

donde la variable dependiente y es métrica y las variables independientes son no métricas. Se trata por tanto de otro caso particular del modelo de regresión múltiple.

El ***análisis de la covarianza simple*** es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas), parte de las cuales son no métricas, siendo la otra parte métricas (*covariables*).

La expresión funcional del modelo del análisis de la covarianza simple ANCOVA es la siguiente:

$$y = F(x_1, x_2, \cdots, x_n)$$

donde la variable dependiente y es métrica y las variables independientes son algunas métricas y otras no métricas. Se trata por tanto de otro caso particular del modelo de regresión múltiple.

El ***análisis de la varianza múltiple*** es una técnica estadística utilizada para analizar la relación entre varias variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) no métricas. El objetivo esencial de los modelos del análisis de la varianza múltiple es contrastar si los valores no métricos de las variables independientes determinarán la igualdad de vectores de medias de una serie de grupos determinados por ellos en las variables dependientes. De modo que el modelo MANOVA mide la significación estadística de las diferencias entre los vectores de medias de los grupos determinados en las variables dependientes por los valores de las variables independientes.

La expresión funcional del modelo del análisis de la varianza múltiple MANOVA es la siguiente:

$$G(y_1, y_2, \cdots, y_m) = F(x_1, x_2, \cdots, x_n)$$

donde las variables dependientes son métricas y las variables independientes son no métricas. Seguimos hablando de una caso particular de la regresión múltiple.

El ***análisis de la covarianza múltiple*** es una técnica estadística utilizada para analizar la relación entre varias variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) mezcla de variables métricas y no métricas.

La expresión funcional del modelo del análisis de la covarianza múltiple MANCOVA es la siguiente:

$$G(y_1, y_2, \cdots, y_m) = F(x_1, x_2, \cdots, x_n)$$

donde las variables dependientes son métricas y las variables independientes son una parte métricas y otra parte no métricas.

En el análisis de la covarianza, tanto simple como múltiple, las variables métricas independientes (*covariables*) tienen como objetivo eliminar determinados efectos que puedan sesgar los resultados incrementando la varianza dentro de los grupos. En el análisis de la covarianza se suele comenzar eliminando, mediante una regresión lineal, la variación experimentada por las variables dependientes producida por la covariable o covariables de efectos indeseados, para continuar con un análisis ANOVA o MANOVA sobre las variables dependientes ajustadas (residuos de la regresión anterior).

La regresión múltiple admite la posibilidad de trabajar con variables independientes no métricas si se emplean variables ficticias para su transformación en métricas. A cada clase de la variable no métrica se le asigna un valor numérico.

El ***modelo de regresión múltiple con variables ficticias*** es similar al análisis de la regresión múltiple con la diferencia de que las variables independientes pueden ser también no métricas. Por tanto, se trata de una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas) métricas, no métricas o mezcla de ambas. El objetivo esencial del análisis de la regresión múltiple es utilizar las variables independientes, cuyos valores son conocidos, para predecir la única variable criterio (dependiente) seleccionada por el investigador.

La expresión funcional del análisis de la regresión múltiple con variables ficticias es la siguiente:

$$y = F(x_1, x_2, \cdots, x_n)$$

Al igual que la regresión múltiple, los modelos de elección discreta admiten la posibilidad de trabajar con variables independientes no métricas si se emplean variables ficticias para su transformación en métricas.

Realmente es muy interesante observar que todas las técnicas predictivas para la modelización expuestas hasta aquí son un caso particular o una extensión del modelo de regresión múltiple.

Podríamos tabular los *métodos del análisis multivariante de la dependencia, según la naturaleza de sus variables dependientes e independientes*, como sigue:

TÉCNICA	*Variables dependientes*	*Variables independientes*
ANOVA y MANOVA	Métrica (métricas)	No métricas
ANCOVA y MANCOVA	Métrica (métricas)	Métricas y no métricas
REGRESIÓN MÚLTIPLE	Métrica	Métricas
REGRESIÓN MÚLTIPLE (VARIABLES FICTICIAS)	Métrica	Métricas y no métricas
ECUACIONES SIMULTÁNEAS	Métricas y no métricas	Métricas y no métricas
ELECCIÓN DISCRETA	No métrica	Métricas
ELECCIÓN DISCRETA (VARIABLES FICTICIAS)	No métrica	Métricas y no métricas

CAPÍTULO 1

MODELO LINEAL DE REGRESIÓN MÚLTIPLE. HIPÓTESIS, ESTIMACIÓN, INFERENCIA Y PREDICCIÓN

MODELO LINEAL DE REGRESIÓN MÚLTIPLE

El modelo de regresión múltiple tiene como objetivo explicar el comportamiento de una variable endógena, explicada o dependiente, que designaremos como Y, utilizando la información proporcionada por los valores tomados por un conjunto de variables explicativas, exógenas o independientes, que designaremos por $X_1, X_2, ..., X_k$.

El modelo lineal (modelo econométrico) viene dado de la forma:

$$Y=\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + u$$

Los coeficientes (parámetros) $\beta_1, \beta_2, ..., \beta_k$ denotan la magnitud del efecto que las variables explicativas (exógenas o independientes) $X_1, X_2, ..., X_k$ tienen sobre la variable explicada (endógena o dependiente) Y. El coeficiente β_0 se denomina término constante (o independiente) del modelo. El término ε se denomina término de error del modelo.

Disponemos de un conjunto de T observaciones para cada una de las variables endógena y exógenas, una por cada período temporal. Entonces, podremos escribir el modelo de la forma:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t \quad t = 1,2,3, ...,T$$

La aparición (no necesaria) de un término independiente en el modelo puede interpretarse como la presencia de una primera variable X_0 cuyo valor sea siempre 1.

El problema fundamental que se aborda es el siguiente: suponiendo que la relación entre la variable Y y el conjunto de variables $X_1, X_2, \ldots, X_k$ es como se ha descrito en el modelo, y que se dispone de un conjunto de T observaciones en el tiempo para cada una de las variables, la endógena y las exógenas, ¿cómo pueden asignarse valores numéricos a los parámetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ basándonos en la información muestral? Estos valores se llamarán estimaciones de los parámetros.

Una vez encontradas las estimaciones de los parámetros del modelo, podremos hacer predicciones acerca del comportamiento futuro de la variable Y.

HIPÓTESIS EN EL MODELO LINEAL

Las hipótesis básicas que debe cumplir cualquier modelo lineal pueden clasificarse en cuatro grandes grupos según las componentes del propio modelo. El primer grupo de hipótesis lo constituyen las relativas a la perturbación aleatoria u. El segundo grupo de hipótesis lo forman las relativas a los regresores. Un tercer grupo de hipótesis lo forman las relativas a la forma funcional. Un último grupo lo forman las hipótesis relativas al vector de parámetros $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$.

Hipótesis relativas a la perturbación aleatoria

Formulamos el modelo lineal bajo las siguientes hipótesis clásicas relativas a la perturbación aleatoria u.

- La variable u (término de error) es una variable aleatoria con esperanza nula y matriz de covarianzas constante y diagonal (matriz escalar). Es decir que, para todo instante temporal t, la variable ε_t tiene media cero y varianza σ^2 no dependiente de t, y además $Cov(u_i, u_j) = 0$ para todo i y para todo j instantes temporales distintos entre sí. El hecho de que la varianza de ε_t sea constante para todo t (que no dependa de t), se denomina hipótesis de *homoscedasticidad* y puede también expresarse como $V(u|X_1,X_2, \ldots, X_k) = \sigma^2$ y $V(Y|X_1,X_2, \ldots, X_k) = \sigma^2$. El hecho de que $Cov(u_i, u_j) = 0$ para todo instante temporal i distinto de j se denomina hipótesis de *no autocorrelación*.

- El término de error u es una variable aleatoria no observable, lo que implica que la variable Y es aleatoria, ya que depende de la variable aleatoria u.

- También se considera la hipótesis de *normalidad* de los residuos, consistente en que las variables ε_t sean normales para todo t. Dicho de otro modo, el vector de perturbaciones aleatorias del modelo tiene una distribución normal multivariante de media cero $E(\boldsymbol{u}) = \mathbf{0}$ y matriz de varianzas covarianzas escalar $E(\boldsymbol{uu}') = \sigma^2\mathbf{I}$. Podemos escribir $\boldsymbol{u} \rightarrow N(\mathbf{0}, \sigma^2\mathbf{I})$.

Hipótesis relativas a los regresores

Se tendrán en cuenta las siguientes hipótesis relativas a los regresores (variables independientes o exógenas) del modelo:

- Las variables $X_1, X_2, \ldots, X_k$ son linealmente independientes, es decir, no existe relación lineal exacta entre ellas. Esta hipótesis se denomina hipótesis de *independencia*, y cuando no se cumple, decimos que el modelo presenta *multicolinealidad*. Es equivalente decir que la matriz de regresores tiene rango k.

- Las variables $X_1, X_2, \ldots, X_k$ son deterministas (no son variables aleatorias), ya que su valor es constante proveniente de una muestra tomada en el tiempo, y además no están correlacionadas con el término de error u, es decir, $E(u|X_1,X_2, \ldots, X_k) = 0$ (hipótesis de *exogeneidad*).

- Los regresores no tienen errores de observación o de medida.

Hipótesis relativas al vector de parámetros

- El vector de parámetros β es un vector fijo.

Esta hipótesis asegura la estabilidad en el tiempo de las estimaciones. El estudio profundo de esta hipótesis nos llevará a la teoría de la cointegración, que es un tema que se tratará posteriormente.

Hipótesis relativas a la forma funcional

- La relación entre Y y $X_1, X_2, \ldots, X_k$ es efectivamente lineal (hipótesis de *linealidad*).

- También se supone la ausencia de errores de especificación, es decir, que suponemos que todas las variables X que son relevantes para la explicación de la variable Y, están incluidas en la definición del modelo lineal.

ESTIMACIÓN DEL MODELO LINEAL POR MÍNIMOS CUADRADOS ORDINARIOS MCO

Supongamos que queremos ajustar el modelo de regresión lineal múltiple siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

Disponemos de un conjunto de T observaciones para cada una de las variables endógena y exógenas (una para cada instante temporal). Entonces, podremos escribir el modelo de la forma:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + u_t \quad t = 1,2,3, \ldots, T$$

El criterio de mínimos cuadrados ordinarios MCO considera que la función que mejor se ajusta a los datos es la que minimiza la varianza del error u, lo que es equivalente a minimizar:

$$S(\beta_0, \beta_1, \cdots, \beta_k) = \sum_{t=1}^{T} u_t^2 = \sum_{t=1}^{T} (y_t - (\beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt}))^2$$

Derivando respecto de los parámetros $\beta_0, \beta_1, \ldots, \beta_k$, e igualando a cero tenemos:

$$\left.\begin{array}{l} \dfrac{\partial S}{\partial \beta_0} = 2\sum_{t=1}^{T} (y_t - (\beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \phi_k x_{kt}))(-1) = 0 \\ \dfrac{\partial S}{\partial \beta_1} = 2\sum_{t=1}^{T} (y_t - (\beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt}))(-x_{1t}) = 0 \\ \vdots \\ \dfrac{\partial S}{\partial \beta_k} = 2\sum_{t=1}^{T} (y_t - (\beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt}))(-x_{kt}) = 0 \end{array}\right\} \Rightarrow$$

$$\Rightarrow \left\{\begin{array}{l} \sum_{t=1}^{T} y_t = T\beta_0 + \beta_1 \sum_{t=1}^{T} x_{1t} + \cdots + \beta_k \sum_{t=1}^{T} x_{kt} \\ \sum_{t=1}^{T} y_t x_{1t} = \beta_0 \sum_{t=1}^{T} x_{1t} + \beta_1 \sum_{t=1}^{T} x_{1t}^2 + \cdots + \beta_k \sum_{t=1}^{T} x_{1t} x_{kt} \\ \vdots \\ \sum_{t=1}^{T} y_t x_{kt} = \beta_0 \sum_{t=1}^{T} x_{kt} + \beta_1 \sum_{t=1}^{T} x_{kt} x_{1t} + \cdots + \beta_k \sum_{t=1}^{T} x_{kt}^2 \end{array}\right.$$

Estas ecuaciones forman un sistema denominado sistema de ecuaciones normales, que puede resolverse para β_0, β_1, ..., β_k mediante cualquier método apropiado para resolver sistemas de ecuaciones lineales. De esta forma se obtiene la estimación del modelo.

Alternativamente, podemos partir de la notación matricial del modelo $y = X\beta + u$, donde y es el vector $T \times 1$, X es una matriz $T \times k$, β es un vector $k \times 1$ y u es un vector $T \times 1$. Denotando el modelo estimado por $\hat{y} = X\hat{\beta}$, el vector de residuos se puede expresar, por tanto, como $\hat{u} = y - \hat{y} = y - X\hat{\beta}$. Si designamos con S a la suma de los cuadrados de los residuos, tendremos:

$$S = \hat{u}'\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_T) \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_T \end{bmatrix} = \sum_{t=1}^{T} \hat{u}_t^2$$

Que también puede escribirse como:

$$S = (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} = y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$$

ya que $\hat{\beta}'X'y = y'X\hat{\beta}$ (se trata del mismo escalar en ambas expresiones).

Aplicar el criterio de mínimos cuadrados es equivalente a minimizar el escalar S. Para minimizar S se calcula la primera derivada de S con respecto al vector de coeficientes mínimo cuadráticos, $\hat{\beta}$, y se tiene:

$$\frac{\delta S}{\delta \hat{\beta}} = -2X'y + 2X'X\hat{\beta}$$

Al igualar a cero se obtiene el sistema de ecuaciones normales:

$$X'X\hat{\beta} = X'y$$

Para poder resolver el sistema respecto a $\hat{\beta}$ unívocamente, se debe cumplir que el rango de la matriz $X'X$ sea igual a k. Si se cumple esta condición, se pueden premultiplicar ambos miembros del sistema por $[X'X]^{-1}$:

$$[X'X]^{-1}[X'X]\hat{\beta} = [X'X]^{-1}X'y \Rightarrow \hat{\beta} = [X'X]^{-1}X'y$$

Ya tenemos la expresión del vector de estimadores mínimo cuadráticos:

$$\hat{\beta} = [X'X]^{-1} X'y$$

Además, S presenta un mínimo en $\hat{\beta}$, ya que la matriz de segundas derivadas, $2X'X$, es definida positiva.

ESTIMACIÓN DEL MODELO LINEAL POR MÁXIMA VEROSIMILITUD

Deseamos estimar por máxima verosimilitud β y σ^2 del modelo $y = X\beta + u$. Para ello tomaremos como estimadores aquellos valores que maximizan la densidad de probabilidad conjunta para una muestra de T observaciones, utilizando las hipótesis básicas $u \to N(0, \sigma^2 I)$ e $y \to N(X\beta, \sigma^2 I)$.

La densidad de probabilidad de y o la función de verosimilitud, considerando X e y fijos y β y σ^2 variables:

$$L = \frac{1}{(2\pi\sigma^2)^{T/2}} e^{-(1/2\sigma^2)(y-X\beta)'(y-X\beta)}$$

Como el máximo para L se alcanza en el mismo punto que para lnL, por ser la función logaritmo monótona, podemos, a efectos de maximización, trabajar con lnL en vez de L. Entonces:

$$\ln L = -\frac{T\ln(2\pi)}{2} - \frac{T\ln\sigma^2}{2} - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

Para maximizar $ln\ L$ derivamos respecto a β y σ^2:

$$\frac{\delta \ln L}{\delta\beta} = -\frac{1}{2\sigma^2}(-2X'y + 2X'y + 2X'X\beta),$$

$$\frac{\delta \ln L}{\delta\sigma^2} = -\frac{T}{2\sigma^2} + \frac{(y - X\beta)'(y - X\beta)}{2\sigma^4}.$$

Igualando a cero las derivadas parciales se tienen que el estimador máximo verosímil de β, que denotaremos por $\tilde{\beta}$, cumple que:

$$X'X\tilde{\beta} = X'y$$

Si suponemos que $X'X$ es invertible, tenemos:

$$\tilde{\beta} = [X'X]^{-1} X'y$$

Con lo que tenemos que el estimador máximo-verosímil de β, bajo las hipótesis básicas, coincide con el estimador mínimo-cuadrático. De la derivada parcial respecto a σ^2 se deduce que el estimador máximo verosímil de σ^2 es $\frac{\hat{u}'\hat{u}}{T}$.

INFERENCIA EN EL MODELO POR MÍNIMOS CUADRADOS ORDINARIOS MCO

Ya sabemos que el modelo lineal de regresión múltiple con series temporales puede escribirse de la forma:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t \quad t = 1,2,3, ...,T$$

La expresión anterior puede representarse en forma matricial como sigue:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_t \end{pmatrix} = \begin{pmatrix} 1\ X_{11}\ X_{21} \cdots X_{k1} \\ 1\ X_{12}\ X_{22} \cdots X_{k2} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ 1\ X_{1T}\ X_{2T} \cdots X_{kT} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_t \end{pmatrix}$$

Abreviadamente podemos poner: $Y = X\beta + u$

El primer objetivo del análisis econométrico es el de obtener estimaciones, es decir, valores numéricos de los coeficientes β_0, β_1, β_2, ..., β_k como función de la información muestral. Estas estimaciones pueden ser también por intervalos, es decir, que podremos calcular intervalos de confianza para los parámetros.

Supongamos que disponemos ya de un vector de estimaciones $\hat{B}$ de los coeficientes. Podríamos escribir:

$$\hat{Y} = X\hat{B} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k$$

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_{1t} + \hat{\beta}_2 X_{2t} + ... + \hat{\beta}_k X_{kt} \quad t = 1,2,3,...,T$$

Los *residuos* son, por definición, las diferencias entre los verdaderos valores de la variable Y_t y los valores estimados para Y_t. Es decir, $\hat{u}_t = Y_t - \hat{Y}_t$ para todo t.

De aquí deducimos que $Y = \hat{Y} + \hat{u} = X\hat{B} + \hat{u}$, con lo que el modelo original es $Y = X\beta + u$, y el modelo estimado será $Y = X\hat{B} + \hat{u}$. Las estimaciones de los parámetros pueden calcularse por el método de mínimos cuadrados, consistente en minimizar la suma de los cuadrados de los residuos, también llamada *suma residual* (*SR*), cuya expresión es la siguiente:

$$SR = \hat{u}'\hat{u} = \sum_{t=1}^{T} \hat{u}_t^{\,2} = \sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2$$

Ya hemos visto que el valor de las *estimaciones de los parámetros (por mínimos cuadrados o por máxima verosimilitud)* viene dado por la expresión $\hat{B} = (X'X)^{-1}X'Y$. Dichas estimaciones son insesgadas, pues $E(\hat{B}) = B$. La matriz de covarianzas de $\hat{B}$ resulta ser $\sigma^2 (X'X)^{-1}$. Para comprobarlo tenemos:

$$\hat{\beta} = [X'X]^{-1}X'[X\beta + u] = [X'X]^{-1}X'[X\beta + u] = \beta + [X'X]^{-1}X'u$$

Aplicando esperanzas y teniendo en cuenta que β es un vector fijo y X una matriz fija, se tiene:

$$E(\hat{\beta}) = \beta + [X'X]^{-1}X'E(u) = \beta .$$

La matriz de varianzas-covarianzas del vector de estimadores mínimo-cuadráticos viene dada por:

$$E[\hat{\beta} - (\hat{\beta})][\hat{\beta} - (\hat{\beta})]' = E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] = E\left[[X'X]^{-1}X'uu'X[X'X]^{-1}\right] =$$
$$= [X'X]^{-1}X'E[uu']X[X'X]^{-1} = [X'X]^{-1}X'[\sigma^2 I]X[X'X]^{-1} = \sigma^2[X'X]^{-1}$$

La varianza de $\hat{\beta}_i$ vendrá entonces dada por $\sigma^2 a_{ii}$, donde a_{ii} es el elemento *ii*-ésimo de $[X'X]^{-1}$. Análogamente, la covarianza entre $\hat{\beta}_i$ y $\hat{\beta}_j$ será $\sigma^2 a_{ij}$, siendo a_{ij} el elemento *ij*-ésimo de $[X'X]^{-1}$.

De los resultados anteriores deducimos que el estimador $\hat{\beta}_i$ de uno cualquiera de los coeficientes β_i, tiene como esperanza matemática β_i, y como desviación típica el valor $\sigma^2 a_{ii}$, donde a_{ii} es el elemento *i*-ésimo en la diagonal principal de la matriz $\sigma^2 (X'X)^{-1}$.

Luego, bajo la hipótesis de normalidad de los residuos, el estadístico:

$$N_i = \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{a_{ii}}}$$

sigue una distribución normal (0,1).

El estimador (máximo verosímil y de mínimos cuadrados) de σ^2 es $\dfrac{\hat{u}'\hat{u}}{T}$, pero este estimador no es insesgado. Un *estimador insesgado de la varianza del error* es:

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{T - k - 1}$$

Por otra parte, también se demuestra que el estadístico $G = u'u/\sigma^2$ sigue una distribución Chi-cuadrado con $T - k - 1$ grados de libertad, lo cual nos va a permitir calcular *intervalos de confianza y contrastes de hipótesis para σ y para su cuadrado.*

Las distribuciones de los estadísticos N_i y G nos llevan a la conclusión de que el estadístico $N_i / [G/(T-k-1)]^{1/2}$ es una t de Student con $T - k - 1$ grados de libertad, lo que permite afirmar que el estadístico:

$$T_i = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{a_{ii}}}$$

sigue una distribución t de Student con $T-k-1$ grados de libertad, lo cual nos va a permitir *hallar intervalos de confianza y contrastes de hipótesis para los parámetros β_i del modelo.* Se podrá contrastar la hipótesis nula H_o de que $\beta_i = 0$ para cada $i=1,2,\ldots,T$ de la forma habitual utilizando el estadístico T_i; es decir, si T_{i0} es el valor de T_i cuando $\beta_i = 0$, se aceptará la hipótesis *Ho* al nivel α cuando $T_{i0} \leq t_{\alpha/2,T-k-1}$. El intervalo de confianza para β_i al nivel α vendrá dado por $\beta_i \pm t_{\alpha/2,T-k-1}\ \hat{\sigma}\sqrt{a_{ii}}$ ($t_{\alpha/2,T-k-1}$ es el valor de la abscisa de una t de Student con $T-k-1$ grados de libertad, que deja a su derecha $\alpha/2$ de área).

Introducimos ahora los siguientes conceptos:

- *Suma total* $ST = \sum_{i=1}^{T} (Y_t - \bar{Y})^2 = Y'Y - T\bar{Y}^2$.
- *Suma explicada* $SE = \sum_{i=1}^{T} (\hat{Y}_t - \bar{Y})^2 = \hat{Y}'Y - T\bar{Y}^2$.
- *Suma residual* $SR = \sum_{i=1}^{T} (Y_t - \hat{Y})^2 = \hat{u}'\hat{u}$ (concepto ya conocido).

La *suma total* es la varianza muestral de la variable endógena (salvo el factor tamaño muestral), y es por tanto una medida del tamaño de las fluctuaciones experimentadas por dicha variable alrededor de su valor medio. El objeto fundamental de todo modelo econométrico es tratar de explicar dichas fluctuaciones.

La *suma explicada* es el grado de fluctuación de la variable $\hat{Y}_t$ alrededor del promedio de Y. Por tanto, la *suma explicada* es el nivel de fluctuación de la variable Y_t que el modelo es capaz de explicar. Es la variación explicada por los regresores.

La *suma residual*, ya introducida previamente, es un indicador del nivel de error del modelo en su intento de explicar la evolución temporal de la variable Y_t.

Ya sabemos que:

$$SR = \sum_{i=1}^{T}(Y_t - \hat{Y})^2 = \hat{u}'\hat{u} = (Y - X\hat{B})'(Y - X\hat{B}) = Y'Y - \hat{B}'X'Y = Y'Y - \hat{Y}'Y$$

Luego podemos escribir la igualdad $Y'Y = \hat{Y}'Y + \hat{u}'\hat{u}$, y si a los dos miembros de esta igualdad les restamos $T\,\overline{Y}^2$, tenemos que:

$(Y'Y - T\overline{Y}^2) = (\hat{Y}'Y - T\overline{Y}^2) + \hat{u}'\hat{u}$, o sea, $ST = SE + SR$.

Luego se tiene: suma total = suma explicada + suma residual

A estos tres términos se les llama *Suma de cuadrados*.

A cada suma de cuadrados dividida por sus grados de libertad se le llama cuadrado medio. Bajo la hipótesis de normalidad de los residuos, SE se distribuye según una Chi-cuadrado con k grados de libertad, SR según una Chi-cuadrado con $T-k-1$ grados de libertad, y ST según una *Chi-cuadrado* con $n-1$ grados de libertad. Por tanto el *Cuadrado Medio explicado por el modelo* será $CM(E) = SE/k$, y el *Cuadrado Medio residual* será $CM(R) = SR/(T-k-1)$.

Se define el *coeficiente de determinación* (R^2) como una medida descriptiva del ajuste global del modelo cuyo valor es el cociente entre la variabilidad explicada (o *suma explicada*) y la variabilidad total (o *suma total*), o sea, $R^2 = SE/ST = 1 - SR/ST$.

Un modelo será tanto mejor cuanto mayor sea R^2, aunque esta afirmación no sea demasiado severa, ya que este coeficiente depende mucho de nuevas variables introducidas en el modelo, aunque éstas no empeoren la calidad de la regresión.

Este problema se arregla sustituyendo este coeficiente por el coeficiente de determinación corregido, que para muestras grandes ya no va a depender del número de variables del modelo.

Se define el *coeficiente de correlación múltiple* como la raíz cuadrada del coeficiente de determinación, y su valor es R.

Se define *el coeficiente de determinación corregido por los grados de libertad (coeficiente de determinación ajustado)* como el valor:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{T-1}{T-k-1}$$

Se observa que cuando $T \to \infty$, o sea, para muestras grandes, $(T-1)/(T-k-1) \to 1$ y no depende de k, que es el número de variables del modelo. Además, $T \to \infty \Rightarrow \overline{R}^2 \to R^2$.

Ahora ya podemos considerar a $\overline{R}^2$ como una buena medida de la calidad de la regresión. El modelo será tanto mejor cuanto mayor sea el coeficiente de determinación corregido $\overline{R}^2$.

De las distribuciones de SE y SR, se deduce que el estadístico:

$$F = \frac{SE/k}{SR/(T-k-1)}$$ tiene una distribución $F(k, T-k-1)$ de Fisher-Snedecor.

Pero como $1-R^2 = SR/ST$, podemos poner F de la forma:

$$F(k, T-k-1) = \frac{R^2}{(1-R^2)}\frac{(T-k-1)}{k}$$

Por tanto el estadístico F nos permitirá hacer *contrastes sobre el coeficiente de correlación*. Para el caso de regresión simple ($k = 1$) tenemos una $F(1, T-2)$ que equivale a una t de Student con $T-2$ grados de libertad.

El estadístico $\dfrac{(\hat{B}-B)'X'X(\hat{B}-B)}{k\hat{\sigma}^2}$ sigue una distribución $F(k, T-k-1)$.

Este estadístico va a permitir hallar *regiones de confianza a un nivel de significación α para el conjunto de parámetros β_i del modelo*. Este estadístico también nos va a permitir contrastar la hipótesis nula $\beta_1=\beta_2=...=\beta_k=0$.

El cuadro del análisis de la varianza quedará como sigue:

Fuente de variación	*Suma de cuadrados*	*Grados de libertad*	*Cuadrados medios*	*F*
Modelo	*SE*	*k*	*CM(E)=SE/k*	$\frac{CM(E)}{CM(R)}$
residual	*SR*	*T-k-1*	*CM(R)=SR/(T-k-1)*	
Total	*ST*	*T-1*		

El estadístico más general:

$$T=\frac{(D\hat{B}-DB)'[D(X'X)^{-1}D']^{-1}(D\hat{B}-DB)}{k\hat{\sigma}^2}$$

también sigue una distribución $F(k,T-k-1)$ para una matriz adecuada D.

Este estadístico va a permitir realizar contrastes más generales de diversas clases de hipótesis y construir regiones de confianza para los parámetros del modelo y para las predicciones. Para ello basta tomar las formas adecuadas de la matriz D.

Para contrastar la hipótesis $\beta_1=\beta_1^*$, $\beta_2=\beta_2^*,\ldots,\beta_k=\beta_k^*$, tomamos:

$$D=\begin{pmatrix}1&0&\cdots&0\\0&1&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&1\end{pmatrix}=I_{kxk}\Leftrightarrow DB=\begin{pmatrix}1&0&\cdots&0\\0&1&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&1\end{pmatrix}\underbrace{\begin{pmatrix}\beta_1\\\beta_2\\\vdots\\\beta_k\end{pmatrix}}_{B}=\underbrace{\begin{pmatrix}\beta_1^*\\\beta_2^*\\\vdots\\\beta_k^*\end{pmatrix}}_{B^*}\Leftrightarrow$$

$$\Leftrightarrow T=\frac{(\hat{B}-B^*)'X'X(\hat{B}-B^*)}{k\hat{\sigma}^2}\rightarrow F_{k,T-k-1}$$

Para contrastar un subconjunto de parámetros $\beta_{r+1}=\beta_{r+1}^*$, $\beta_{r+2}=\beta_{r+2}^*$ $,\ldots,\beta_{r+k}=\beta_{r+k}^*$, tomamos:

$$D=\begin{pmatrix}0&0&\cdots&0&1&0&\cdots&0\\0&0&\cdots&0&0&1&\cdots&0\\\vdots&\vdots&\ddots&\vdots&\vdots&\vdots&\ddots&\vdots\\\underbrace{0\quad 0\quad\cdots\quad 0}_{r}&&&&\underbrace{0\quad 0\quad\cdots\quad 1}_{k-r}&&&\end{pmatrix}=I_{kxk}\Leftrightarrow$$

$$\Leftrightarrow T=\frac{(D\hat{B}-B^*)'[D[X'X]^{-1}D']^{-1}(D\hat{B}-B^*)}{(k-r)\hat{\sigma}^2}\rightarrow F_{k-r,T-k-1}$$

También es posible contrastar un conjunto de restricciones lineales sobre los parámetros, que pueden escribirse en general de la siguiente forma:

$$\begin{aligned}&a_{11}\beta_1+a_{12}\beta_2+\cdots+a_{1k}\beta_k=\beta_1^*\\&a_{21}\beta_1+a_{22}\beta_2+\cdots+a_{2k}\beta_k=\beta_2^*\\&\vdots\\&a_{r1}\beta_1+a_{r2}\beta_2+\cdots+a_{rk}\beta_k=\beta_r^*\end{aligned}$$

tomando:

$$D=\begin{pmatrix}a_{11}&a_{12}&\cdots&a_{1k}\\a_{21}&a_{22}&\cdots&a_{2k}\\\vdots&\vdots&\ddots&\vdots\\a_{r1}&a_{r2}&\cdots&a_{rk}\end{pmatrix}\text{ y }T=\frac{(D\hat{B}-DB)'[D(X'X)^{-1}D'(D\hat{B}-DB)}{r\hat{\sigma}^2}\rightarrow F_{r,T-k-1}$$

El teorema de Gauss-Markov asegura que en el contexto del modelo de regresión lineal (bajo los supuestos típicos) los estimadores MCO de los parámetros β_0, β_1, β_2, ..., β_k son los de menor varianza entre los estimadores lineales e insesgados. Además, los estimadores MCO, $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_k$ son estimadores consistentes de β_0, β_1, β_2, ..., β_k:

$$p\lim\hat{\beta}_j=\beta_j\quad j=0,1,\ldots,k$$

Dicho de otra forma, los estimadores MCO $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_k$ son estimadores consistentes de los parámetros β_0, β_1, β_2, ..., β_k porque su sesgo tiende a cero cuando el tamaño muestral tiende a infinito.

PREDICCIONES

Se puede definir una predicción como un intento de anticipar el futuro. En el contexto temporal, y tratándose de procedimientos cuantitativos, puede hablarse de dos clases de predicciones: condicionales e incondicionales. Las ***predicciones condicionales*** son las que se realizan mediante modelos causales. Por ejemplo, en un modelo de regresión de series temporales que relaciona dos variables, una dependiente, Y, y otra independiente, X, las predicciones de Y están condicionadas a X, es decir, se predice Y dada X.

Las ***predicciones incondicionales*** son las que se hacen mediante métodos autoproyectivos. Estos métodos pueden estar basados en dos enfoques alternativos: el determinista, o clásico, y el estocástico, o moderno (basado en la metodología de Box y Jenkins). El *enfoque determinista* es más adecuado cuando se dispone de un número limitado de observaciones, mientras que el *enfoque estocástico* es más adecuado cuando las series son de mayor tamaño.

Para cada tipo de predicciones (a corto, medio y largo plazo), existen determinados métodos más adecuados. Por ejemplo, el análisis de tendencias es un método para realizar predicciones a largo plazo, los modelos econométricos son adecuados para hacer predicciones a corto y medio plazo, y los métodos autoproyectivos son más adecuados para realizar predicciones a corto plazo.

Como hemos visto, una de las finalidades del análisis de los modelos econométricos es hacer predicciones condicionales para la variable dependiente. Si estimamos el modelo $Y=XB$ y obtenemos el modelo estimado $\hat{Y} = X\hat{B}$, tenemos que el valor $\hat{Y}_0 = X_0\hat{B}$ es un estimador lineal insesgado óptimo del pronóstico de Y, para un valor dado X_0 de X.

Predicciones puntuales y en media

Se puede predecir la media, $E(Y_0)$, o el valor puntual, Y_0. Los errores de predicción vendrán cuantificados por las varianzas de los predictores. La varianza para la predicción en media vale $\sigma^2 X_0(X'X)^{-1} X_0'$.

Un *intervalo de confianza para la predicción en media* al nivel α vendrá dado por la expresión $\hat{Y}_0 \pm t_{T-k-1}(\alpha/2)\hat{\sigma}\sqrt{X_0(X'X)^{-1}X_0'}$, donde $t_{(T-k-1)}(\alpha/2)$ es el valor de la t de Student con $T-k-1$ grados de libertad en el punto $\alpha/2$.

La varianza para la predicción puntual vale $\sigma^2\ (X_0(X'X)^{-1}\ X_0'+1)$, y un *intervalo de confianza para la predicción puntual* vendrá dado por la expresión definida como $\hat{Y}_0 \pm t_{T-k-1}(\alpha/2)\hat{\sigma}\sqrt{1+X_0(X'X)^{-1}X_0'}$, donde $t_{(T-k-1)}(\alpha/2)$ es el valor de la t de Student con $T-k-1$ grados de libertad en el punto $\alpha/2$.

Capacidad predictiva de un modelo

Para evaluar la capacidad predictiva del modelo con Eviews se utilizan varios estadísticos alternativos. Siendo n el horizonte de predicción, los estadísticos más habituales para la evaluación de la capacidad predicativa son los siguientes:

Raíz del error cuadrático medio (Root Mean Squared Error):

$$RECM = \sqrt{\frac{\sum_{i=1^\circ}^{n}(\hat{Y}_i - Y_i)^2}{n}}$$

Error absoluto medio (Mean Absolute Error):

$$EAM = \frac{\sum_{i=1^\circ}^{n}|\hat{Y}_i - Y_i|}{n}$$

Error absoluto medio del porcentaje del error (Mean Abs. Percent Error):

$$EAMP = \frac{1}{n}\sum_{i=1^\circ}^{n}\left|\frac{\hat{Y}_i - Y_i}{Y_i}\right|$$

Coeficiente de desigualdad de Theil (Theil Inequality Coefficient):

$$CDT = \frac{\sqrt{\frac{\sum_{i=1^\circ}^{n}(\hat{Y}_i - Y_i)^2}{n}}}{\sqrt{\frac{\sum_{i=1^\circ}^{n}\hat{Y}_i^2}{n}} + \sqrt{\frac{\sum_{i=1^\circ}^{n}Y_i^2}{n}}}$$

Proporción del sesgo (Bias Proportion):

$$\frac{(\bar{\hat{Y}} - Y)^2}{\sum_{i=1^\circ}^{n}(\hat{Y}_i - Y_i)^2 / n}$$

Proporción de la varianza (Variance Proportion):

$$\frac{(S_{\hat{Y}} - S_Y)^2}{\sum_{i=1^\circ}^{n}(\hat{Y}_i - Y_i)^2 / n}$$

Proporción de la covarianza (Covariance Proportion):

$$\frac{2(1-r)S_{\hat{Y}}S_Y)}{\sum_{i=1^{\circ}}^{n}(\hat{Y}_i - Y_i)^2 / n}$$

Cuanto más próximos estén a cero los valores de los cuatro primeros estadísticos, mejor será la capacidad predictiva del modelo, lo que permitirá comparar un modelo con otros alternativos. Las tres proporciones varían entre cero y uno y también es conveniente que sean pequeñas.

SELECCIÓN DE MODELOS DE REGRESIÓN

Cuando se ajusta un modelo de regresión múltiple podemos tener problemas de diferentes tipos. Ante esta situación existen criterios que permiten elegir el mejor modelo para unas variables y un conjunto de datos dados.

Aparte del *criterio de regresión hacia adelante*, que va incluyendo variables en el modelo hasta obtener el ajuste ideal, el *criterio de regresión hacia atrás* que empieza incluyendo todas las variables en el modelo y va eliminando las adecuadas hasta obtener un ajuste óptimo libre de problemas y el *criterio de selección paso a paso*, hay otros criterios de selección de modelos basados en estadísticos como R^2 y *Cp* de Mallows, que permiten obtener modelos bien ajustados muy útiles en ajustes exploratorios.

Asimismo, los estadísticos AIC de AKAIKE y SC de Schwarz permiten *seleccionar el modelo ajustado con mejor capacidad explicativa* aquel que presenta menor valor para estos estadísticos. La definición de cada uno de ellos es:

$$AIC = -\frac{2l}{T} + \frac{2(K+1)}{T}$$

$$SC = -\frac{2l}{T} + \frac{(K+1)\log(T)}{T}$$

$$l = -\frac{T}{2}(1 + \log(2\pi) + \log\frac{e'e}{T})$$

Donde K es el número de variables independientes del modelo (sin incluir la constante), T es el tamaño muestral o número de observaciones de que se dispone para la estimación del modelo y e es el error del modelo.

ANÁLISIS DE LOS RESIDUOS

Una vez construido el modelo de regresión con series temporales, tendremos que contrastar, entre otras, las hipótesis de linealidad, normalidad, homoscedasticidad, no autocorrelación e independencia, materia que se ampliará en sucesivos capítulos de este libro. Los residuos van a presentar una primera información sobre estas hipótesis.

Si el histograma de frecuencias de los residuos no se ajusta al de una normal, pueden existir valores atípicos. Eliminando los pares (X_i Y_i) que producen los valores atípicos, se puede conseguir *normalidad* en los residuos.

Si graficamos los valores de t contra los valores de $\hat{u}_t$ (o sea, si hacemos la gráfica cuyos puntos son los pares (t, $\hat{u}_t$) y detectamos una tendencia creciente o decreciente en el grafo, puede existir *autocorrelación o correlación serial.*

Si graficamos los valores de $\hat{Y}_t$ contra los valores de $\hat{u}_t$, o sea, si hacemos la gráfica cuyos puntos son los pares ($\hat{Y}_t, \hat{u}_t$) y detectamos una tendencia de cualquier tipo en el grafo, puede existir autocorrelación, ya que habrá correlación entre los residuos. También puede haber en este caso *heteroscedasticidad,* o también falta de *linealidad.*

Si graficamos los valores de $\hat{Y}_t$ contra los valores de $\hat{u}_t^2$, o sea, si se hace la gráfica cuyos puntos son los pares ($\hat{Y}_t, \hat{u}_t^2$) y detectamos una tendencia de cualquier tipo en el grafo, puede existir heteroscedasticidad.

Si graficamos los valores de X_t contra los valores de $\hat{u}_t$, o sea, si se hace la gráfica cuyos puntos son los pares (X_t, $\hat{u}_t$) y detectamos una tendencia creciente o decreciente en el grafo, puede existir autocorrelación, ya que los residuos no estarán incorrelados con las variables explicativas. También puede haber heteroscedasticidad, o falta de linealidad.

Si graficamos los valores de X_t contra los valores de $\hat{u}_t^2$, o sea, si se hace la gráfica cuyos puntos son los pares (X_t, $\hat{u}_t^2$) y detectamos cualquier tendencia en el grafo, puede existir heteroscedasticidad o falta de linealidad (habrá relación entre la varianza del término del error y las variables explicativas).

Estos análisis pueden realizarse también utilizando residuos estandarizados o residuos estudentizados, que suelen ser más efectivos para detectar deficiencias en el modelo.

Los residuos estudentizados, cuya distribución es una t de Student con $T-k-2$ grados de libertad, se usan también para detectar valores atípicos en los residuos (análisis de la normalidad de los residuos o de la mala especificación del modelo).

Un contraste muy importante para detectar la autocorrelación es el contraste de Durbin-Watson, según el cual el estadístico:

$$d = \frac{\sum_{t=2}^{T}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T}\hat{u}_t^{\,2}}$$

nos permite adoptar la regla no demasiado rigurosa de que si d vale 0 hay autocorrelación perfecta positiva; si d se aproxima a 2 no hay autocorrelación, y si d se aproxima a 4 hay autocorrelación perfecta negativa. No obstante, d se encuentra tabulado, y según la franja en la que caiga su valor, se acepta o rechaza la hipótesis de autocorrelación.

Si la matriz $X'X$ tiene determinante cero, no podrá calcularse su inversa (ni podrán hacerse las estimaciones de los parámetros), en cuyo caso hay multicolinealidad. Este problema suele resolverse sustituyendo las variables del modelo por un conjunto grande de sus componentes principales para hacer de nuevo la regresión, pero esta vez con las componentes principales como variables independientes del modelo. La multicolinealidad también puede abordarse mediante la matriz de correlaciones de las variables independientes, ya que las variables con correlación alta pueden ser candidatas a provocar multicolinealidad, problema que puede intentar solucionarse eliminando una de las variables fuertemente correlacionadas, siempre que dicha variable no sea relevante.

MODELO LINEAL CON RESTRICCIONES

Supongamos que se contrastan determinadas hipótesis en un modelo que posteriormente pueden ser fijadas a priori. Para estimar un nuevo modelo similar puede ser interesante hacerlo fijando ya las restricciones derivadas de la hipótesis cierta previamente contrastada. El problema ahora será buscar el estimador $\hat{\beta}_R$ que minimice la suma de cuadrados de los residuos, de igual modo que hicimos en el modelo lineal de regresión múltiple, pero esta vez imponiendo las restricciones $R\beta = r$, es decir, que, a diferencia de lo que allí se hizo, se trata esta vez de resolver un problema de optimización sujeto a restricciones lineales. El Lagrangiano de tal problema será:

$$L = (y - X\beta)'(y - X\beta) - 2\lambda'(R\beta - r)$$

donde λ es un vector de dimensión $q \times 1$ de multiplicadores de Lagrange (tantos como restricciones). Tomando derivadas parciales de este Lagrangiano con respecto a β', así como con respecto a λ, se tiene:

$$\frac{\partial L}{\partial \beta'} = -2X'y + 2X'X\beta - 2R'\lambda$$
$$\frac{\partial L}{\partial \lambda} = -2(R\beta - r)$$

Igualando a cero estas derivadas parciales y resolviendo el sistema de $k + q$ ecuaciones que así se obtienen, resulta:

$$X'X\beta - X'y - R'\lambda = 0$$
$$R\beta_R - r = 0$$

Las soluciones de este sistema de ecuaciones nos llevan al estimador de mínimos cuadrados restringidos (MCR).

Premultiplicando $X'X\beta - X'y - R'\lambda = 0$ por $R(X'X)^{-1}$ se tiene:

$$R\beta - R(X'X)^{-1}X'y - R(X'X)^{-1}R'\lambda = 0$$

Bajo la hipótesis nula $R\beta = r$ se tiene:

$$\lambda = \left[R(X'X)^{-1}R'\right]^{-1}\left(r - R\hat{\beta}\right)$$

donde $\hat{\beta}$ es el estimador MCO habitual, es decir, sin imponer ninguna de las restricciones contenidas en Ho. Sustituyendo λ en $X'X\beta - X'y - R'\lambda = 0$ y llamando $\hat{\beta}_R$ al vector de parámetros estimados con restricciones, se tiene:

$$X'X\hat{\beta}_R - X'y - R\left[R(X'X)^{-1}R'\right]^{-1}\left(r - R\hat{\beta}\right) = 0$$

Multiplicando la expresión anterior a la izquierda por $(X'X)^{-1}$ se tiene:

$$(X'X)^{-1}X'X\hat{\beta}_R-(X'X)^{-1}X'y-(X'X)^{-1}R\left[R(X'X)^{-1}R'\right]^{-1}\left(r-R\hat{\beta}\right)=0$$

$$\hat{\beta}_R-\hat{\beta}-(X'X)^{-1}R\left[R(X'X)^{-1}R'\right]^{-1}\left(r-R\hat{\beta}\right)=0$$

$$\hat{\beta}_R=\hat{\beta}+(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}\left(r-R\hat{\beta}\right)$$

Ya hemos calculado el estimador de mínimos cuadrados restringidos $\hat{\beta}_R$ del modelo (MCR). Podemos interpretar esta expresión diciendo que el estimador MCO restringido es una corrección del estimador sin restringir, siendo el tamaño de dicha corrección el segundo término en la expresión de $\hat{\beta}_R$. La corrección será tanto mayor cuanto más lejos esté el estimador no restringido de satisfacer las restricciones.

El estimador MCR es insesgado sólo si las restricciones $R\beta = r$ bajo las que se ha obtenido son ciertas. El estimador MCR difiere del estimador MCO sólo si este último no satisface las restricciones en Ho (lo que en general ocurrirá). Pero, si resulta que el estimador MCO satisface exactamente las restricciones cuya validez se contrasta, entonces el estimador MCR coincide con el estimador MCO. La matriz de covarianzas del estimador MCR es siempre inferior a la matriz de covarianzas del estimador MCO, incluso si las restricciones no son ciertas. Aun pareciendo paradójico, este resultado tiene sentido, pues, al imponer las restricciones, limitamos la región del espacio paramétrico en la que buscamos el estimador mínimo-cuadrático, por lo que podremos estimarlo con una mayor precisión.

La matriz de covarianzas del estimador restringido es:

$$Var\left(\hat{\beta}_R\right)=\sigma_u^2\left[(X'X)^{-1}-(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}\right]$$

Por tanto:

$$Var\left(\hat{\beta}_R\right)-Var\left(\hat{\beta}\right)=\sigma_u^2(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}$$

Esta última expresión es una matriz definida positiva, lo que lleva a que los elementos de la diagonal de la matriz de covarianzas de $\hat{\beta}_R$ sean inferiores a los elementos correspondientes de la matriz de covarianzas de $\hat{\beta}$, para una estimación dada del parámetro σ_u^2.

REGRESIÓN CON VARIABLES CUALITATIVAS: VARIABLES FICTICIAS

Modelos de regresión con variables cualitativas

En los modelos de regresión pueden estar presentes tanto variables cuantitativas como cualitativas. Consideremos en primer lugar los ***modelos con variables explicativas cualitativas***. Si una variable explicativa cualitativa D de un modelo tiene m categorías, se pueden incluir en su lugar en el modelo m-1 ***variables dicotómicas ficticias*** D_i que recogen los efectos de la variable cualitativa sobre las observaciones muestrales. A la categoría a la que no se asigna variable dicotómica se denomina categoría base, de comparación, de control, de referencia u omitida. Dadas las T observaciones de la muestra para ajustar el modelo, tenemos:

$$D_i=\begin{cases}1 \text{ si la observación pertenece a la categoría } i\\ 0 \text{ si la observación no pertenece a la categoría } i\end{cases}\qquad i=1,\cdots,m-1$$

El modelo de regresión se ajustará en la forma habitual tratando a estas variables D_i como nuevos regresores. La razón de introducir m-1 variables ficticias en lugar de m es evitar el problema de la colinealidad perfecta en el modelo, ya que si se introducen m variables ficticias la matriz de datos del modelo será la matriz X que se especifica a continuación en la que la suma de las columnas relativas a las m variables ficticias dé como resultado la columna de unos de la matriz relativa a la constante del modelo. De esta forma, se tiene que una columna de la matriz X se obtiene como combinación lineal exacta de otras columnas. Estamos entonces ante el problema de la presencia de multicolinealidad perfecta en el modelo. Este problema de la multicolinealidad perfecta también puede solucionarse introduciendo las m variables dicotómicas ficticias relativas a las m categorías de la variable cualitativa del modelo, pero considerando el mismo sin término constante. En el modelo relativo a la matriz X se han considerado k variables cuantitativas y una variable cualitativa representada por m variables ficticias.

$$X=\begin{bmatrix}1&1&0&\cdots&0&x_{11}&\cdots&x_{k1}\\ \vdots&\vdots&\vdots&&\vdots&\vdots&&\vdots\\ \vdots&1&0&\cdots&0&\vdots&&\vdots\\ \vdots&0&1&\cdots&0&\vdots&&\vdots\\ \vdots&\vdots&\vdots&&\vdots&\vdots&&\vdots\\ \vdots&\vdots&1&\cdots&0&\vdots&&\vdots\\ \vdots&\vdots&0&\cdots&0&\vdots&&\vdots\\ \vdots&\vdots&\vdots&&\vdots&\vdots&&\vdots\\ \vdots&\vdots&\vdots&&1&\vdots&&\vdots\\ \vdots&\vdots&\vdots&&\vdots&\vdots&&\vdots\\ 1&0&0&\cdots&1&x_{1T}&\cdots&x_{kt}\end{bmatrix}$$

Por ejemplo, si se quiere explicar el salario Y de los profesores de los Institutos españoles mediante la Comunidad Autónoma a la que pertenecen y el gasto en inversión en Educación X de la correspondiente Comunidad de entre las 17, podríamos considerar un modelo con 16 variables dicotómicas D_i y la variable X como variables explicativas. El modelo sería:

$$Y = \alpha_0 + \alpha_1 D_1 + \ldots + \alpha_{16} D_{16} + \alpha_{17} X$$

$$D_i = \begin{cases} 1 \; si \; el \; profesor \; pertenece \; a \; la \; Comunidad \; Autónoma \; i \\ 0 \; si \; el \; profesor \; no \; pertenece \; a \; Comunidad \; Autónoma \; i \end{cases} \qquad i = 1, \cdots, 16$$

También podría considerarse el modelo alternativo:

$$Y = \alpha_1 D_1 + \ldots + \alpha_{17} D_{17} + \alpha_{18} X$$

Las variables dicotómicas ficticias Di relativas a las categorías de la variable cualitativa Comunidad Autónoma de pertenencia de los profesores, miden el efecto relativo de la pertenencia del profesor a la correspondiente Comunidad.

Por otro lado, la presencia ***de interacción entre dos variables ficticias*** puede llevarnos a considerar un término adicional para dicha interacción en el modelo. Por ejemplo, podemos considerar un modelo con dos variables ficticias y una variable cuantitativa como regresores y añadirle el efecto interacción multiplicativo entre las dos variables ficticias.

$$Y = \alpha_0 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_1 D_2 + \alpha_4 X$$

Casos particulares de modelos de regresión con variables explicativas cualitativas son el análisis de la varianza y la covarianza. El *análisis de la varianza simple* es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas) no métricas. El objetivo esencial de los modelos del análisis de la varianza es determinar si diversas muestras proceden de poblaciones con igual media. Los valores no métricos de las variables independientes determinarán una serie de grupos en la variable dependiente. De modo que el modelo ANOVA mide la significación estadística de las diferencias entre las medias de los grupos determinados en la variable dependiente por las categorías de las variables independientes.

El *análisis de la covarianza simple* es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas), parte de las cuales son no métricas, siendo la otra parte métricas (*covariables*).

También podemos considerar los *modelos con variable dependiente cualitativa* (se estudiarán en capítulos posteriores). Los más habituales son los *modelos de elección discreta* en los que se tiene una variable dependiente cualitativa y varias variables independientes cuantitativas. Estos modelos suelen utilizarse para predecir la probabilidad de pertenencia a una categoría (clase) de la variable cualitativa dependiente para valores dados de las variables cuantitativas independientes. Por tanto, los modelos de elección discreta predicen directamente la probabilidad de ocurrencia de un suceso que viene definido por los valores de las variables independientes. Como los valores de una probabilidad están entre cero y uno, las predicciones realizadas con los modelos de elección discreta deben estar acotadas para que caigan en el rango entre cero y uno. El modelo general que cumple esta condición se denomina *modelo lineal de probabilidad.*

Otro tipo de modelos con variable dependiente cualitativa son los *modelos discriminantes* que analizan la relación entre una variable dependiente (o endógena) no métrica (categórica) y varias variables independientes (o exógenas) métricas, de modo que para valores conocidos de las variables independientes se predice con qué categoría (clase) de la variable dependiente se corresponden.

Variables ficticias en el análisis estacional

Son muy habituales las series de tiempo económicas basadas en información mensual o trimestral que presentan patrones estacionales. Suele ser útil eliminar la componente estacional de las series de tiempo con el fin de poderse concentrar en componentes más importantes como la tendencia. El proceso de eliminar la componente estacional de una serie de tiempo se conoce como ***desestacionalización o ajuste estacional*** y la serie obtenida se denomina serie desestacionalizada. Hay muchos métodos para desestacionalizar una serie temporal entre los que se encuentra el método de las variables ficticias dicotómicas.

Supongamos, por ejemplo, que tenemos una serie temporal Y_t con estacionalidad trimestral. Para desestacionalizarla consideramos el modelo:

$$Y_t = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t} + u_t$$

$$D_{it} = \begin{cases} 1 \; en \; el \; trimestre \; i \\ 0 \; en \; el \; resto \end{cases} \qquad i = 1, \cdots, 4$$

Se observa que en el modelo se omite la constante para evitar la colinealidad perfecta. Para que el efecto estacional esté presente, los parámetros estimados del modelo anterior han de ser significativamente distintos de cero individualmente. Los residuos estimados de la regresión anterior $\hat{u}_t = Y_t - \hat{Y}_t$ serán los valores de la serie desestacionalizada.

Variables ficticias en la regresión por tramos

Supongamos un modelo lineal con variable dependiente Y y con variable explicativa X. Supongamos que la relación que liga a Y y a X no es la misma para valores menores que un valor X^* de X que para valores de X superiores a X^*. En este caso habrá que considerar dos tramos de la variable X ($X<X^*$ y $X>X^*$). Esta regresión se abordará considerando la variable ficticia D que vale 0 para valores de X menores que X^* y que vale uno en caso contrario. La ecuación del modelo a ajustar será la siguiente:

$$Y_t = \alpha_1 + \alpha_2 X_t + \alpha_3 (X_t - X^*) D_t + u_t$$

$$D_t = \begin{cases} 1 \; si \; X_t > X^* \\ 0 \; en \; el \; resto \end{cases}$$

Suponiendo $E(u_t) = 0$, se tiene que:

$E(Y_t \,|D_t = 0, X_t, X^*) = \alpha_1 + \alpha_2 X_t$
$E(Y_t \,|D_t = 1, X_t, X^*) = \alpha_1 - \alpha_3 X^* + (\alpha_2 + \alpha_3) X_t$

Se observa que α_2 corresponde a la pendiente de la recta de regresión en el primer tramo ($X<X^*$) y $\alpha_2+\alpha_3$ es la pendiente de la recta de regresión en el segundo tramo ($X>X^*$).

Ejercicio 1-1. Se trata de estudiar si la humedad (x1), temperatura del aire (x2) y presión barométrica (x3) influyen en la emisión de óxido nitroso (y) de los motores diesel de los automóviles. Para ello se realizan mediciones de las variables anteriores en diferentes momentos obteniéndose los siguientes resultados:

y	x_1	x_2	x_3	y	x_1	x_2	x_3
0,90	72,4	76,3	29,18	1,07	23,2	76,8	29,38
0,91	41,6	70,3	29,35	0,94	47,4	86,6	29,35
0,96	34,3	77,1	29,24	1,10	31,5	76,9	29,63
0,89	35,1	68,0	29,27	1,10	10,6	86,3	29,56
1,00	10,7	79,0	29,78	1,10	11,2	86,0	29,48
1,10	12,9	67,4	29,39	0,91	73,3	76,3	29,40
1,15	8,3	66,8	29,69	0,87	75,5	77,9	29,28
1,03	20,1	76,9	29,48	0,78	96,6	78,7	29,29
0,77	72,2	77,7	29,09	0,82	107,4	86,8	29,03
1,07	24,0	67,7	29,60	0,95	54,9	70,9	29,37

Ajustar los datos a un modelo de regresión lineal múltiple que explique las emisiones de óxido nitroso en función de la humedad, la temperatura del aire y la presión barométrica. Como caso particular, cuantificar la emisión de óxido nitroso para una humedad del 50%, una temperatura de 76 grados y una presión barométrica de 29,30.

Se trata de ajustar el modelo lineal de regresión múltiple siguiente:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Comenzamos planteando el sistema de ecuaciones normales:

$$\begin{aligned}
&\sum_{t=1}^{T} y_t = n b_0 + b_1 \sum_{t=1}^{T} x_{1t} + \cdots + b_k \sum_{t=1}^{T} x_{kt} \\
&\sum_{t=1}^{T} y_t x_{1t} = b_0 \sum_{t=1}^{T} x_{1t} + b_1 \sum_{t=1}^{T} x_{1t}^2 + \cdots + b_k \sum_{t=1}^{T} x_{1t} x_{kt} \\
&\vdots \\
&\sum_{t=1}^{T} y_t x_{kt} = b_0 \sum_{t=1}^{T} x_{kt} + b_1 \sum_{t=1}^{T} x_{kt} x_{1t} + \cdots + b_k \sum_{t=1}^{T} x_{kt}^2
\end{aligned}$$

Para los datos de nuestro problema el sistema de ecuaciones normales es el siguiente:

$$\left.\begin{aligned}
20b_0 + 863{,}1b_1 + 1530{,}4b_2 + 587{,}84b_3 &= 19{,}42 \\
863{,}1b_0 + 54876{,}89b_1 + 67000{,}09b_2 + 25283{,}395b_3 &= 779{,}477 \\
1530{,}4b_0 + 67000{,}09b_1 + 117912{,}32b_2 + 44976{,}867b_3 &= 1483{,}437 \\
587{,}84b_0 + 25283{,}395b_1 + 44976{,}867b_2 + 17278{,}5086b_3 &= 571{,}1219
\end{aligned}\right\}$$

La solución de este sistema de ecuaciones es la siguiente:

$$b_0 = -3{,}507778, \quad b_1 = -0{,}002625, \quad b_2 = 0{,}000799, \quad b_3 = 0{,}154155$$

Por tanto, el modelo de regresión lineal múltiple ajustado será:

$$\hat{y} = -3{,}507778 - 0{,}002625x_1 + 0{,}000799x_2 + 0{,}154155x_3$$

Como caso particular, la emisión de óxido nitroso predicha para una humedad del 50%, una temperatura de 76 grados y una presión barométrica de 29,30 será:

$$\hat{y} = -3{,}507778 - 0{,}002625(50) + 0{,}000799(76) + 0{,}154155(29{,}3) = 0{,}9384$$

Ejercicio 1-2. En un estudio para aumentar la supervivencia (variable y en tanto por ciento) de determinada clase de plantas se utilizan concentraciones de tres productos diferentes (variables x1, x2 y x3 en gramos). Para distintos valores de las variables x1, x2 y x3 se obtienen los siguientes porcentajes de aumento de la supervivencia:

y	x_1	x_2	x_3
25,5	1,74	5,30	10,80
31,2	6,32	5,42	9,40
25,9	6,22	8,41	7,20
38,4	10,52	4,63	8,50
18,4	1,19	11,60	9,40
26,7	1,22	5,85	9,90
26,4	4,10	6,62	8,00
25,9	6,32	8,72	9,10
32,0	4,08	4,42	8,70
25,2	4,15	7,60	9,20
39,7	10,15	4,83	9,40
35,7	1,72	3,12	7,60
26,5	1,70	5,30	8,20

Ajustar los datos a un modelo de regresión lineal múltiple que explique el porcentaje de aumento de la supervivencia en función de las concentraciones de los tres productos que inciden sobre la supervivencia de las plantas. Realizar la estimación de la varianza. Construir un intervalo de confianza al 95% para la respuesta media cuando x1 = 3, x2 = 8 y x3 = 9. Construir también un intervalo de predicción al 95% para una respuesta individual del porcentaje de supervivencia cuando x1 = 3, x2 = 8 y x3 = 9. Realizar el contraste de hipótesis b2 = -2,5 contra b2 > -2,5 para α = 0,05. Calcular el coeficiente de determinación R2 y realizar también el contraste de significación conjunta de la regresión y el contraste de significatividad individual parámetro a parámetro.

Los parámetros estimados serán:

$$\hat{B}=(X'X)^{-1}X'Y=\begin{pmatrix} 8{,}0648 & -0{,}0826 & -0{,}0942 & -0{,}7905 \\ -0{,}0826 & 0{,}0085 & 0{,}0017 & 0{,}0037 \\ -0{,}0942 & 0{,}0017 & 0{,}0166 & -0{,}0021 \\ -0{,}7905 & 0{,}0037 & -0{,}0021 & 0{,}0886 \end{pmatrix}\begin{pmatrix} 377{,}5 \\ 1877{,}567 \\ 2246{,}661 \\ 3337{,}78 \end{pmatrix}=\begin{pmatrix} 39{,}1574 \\ 1{,}0161 \\ -1{,}8616 \\ -0{,}3433 \end{pmatrix}$$

Por tanto tenemos:

$$b_0=39{,}1574,\quad b_1=1{,}0161,\quad b_2=-1{,}8616,\quad b_3=-0{,}3433$$

Los cálculos anteriores los hemos llevado a cabo considerando las matrices siguientes:

$$X=\begin{pmatrix} 1 & 1{,}74 & 5{,}30 & 10{,}80 \\ 1 & 6{,}32 & 5{,}42 & 9{,}40 \\ 1 & 6{,}22 & 8{,}41 & 7{,}20 \\ 1 & 10{,}52 & 4{,}63 & 8{,}50 \\ 1 & 1{,}19 & 11{,}60 & 9{,}40 \\ 1 & 1{,}22 & 5{,}85 & 9{,}90 \\ 1 & 4{,}10 & 6{,}62 & 8{,}00 \\ 1 & 6{,}32 & 8{,}72 & 9{,}10 \\ 1 & 4{,}08 & 4{,}42 & 8{,}70 \\ 1 & 4{,}15 & 7{,}60 & 9{,}20 \\ 1 & 10{,}15 & 4{,}83 & 9{,}40 \\ 1 & 1{,}72 & 3{,}12 & 7{,}60 \\ 1 & 1{,}70 & 5{,}30 & 8{,}20 \end{pmatrix}\qquad Y=\begin{pmatrix} 25{,}5 \\ 31{,}2 \\ 25{,}9 \\ 38{,}4 \\ 18{,}4 \\ 26{,}7 \\ 26{,}4 \\ 25{,}9 \\ 32{,}0 \\ 25{,}2 \\ 39{,}7 \\ 35{,}7 \\ 26{,}5 \end{pmatrix}$$

Por tanto, el modelo de regresión lineal múltiple ajustado será:

$$\hat{y}=39{,}1574+1{,}061x_1-1{,}8616x_2-0{,}3433x_3$$

El estimador de la varianza residual vendrá dado por la expresión:

$$\hat{\sigma}^2=\frac{\hat{u}'\hat{u}}{T-k-1}=\frac{(Y-X\hat{B})'(Y-X\hat{B})}{13-3-1}=\frac{Y'Y-(X\hat{B})'Y}{9}=4{,}298$$

No olvidemos que las matrices Y, X y $\hat{B}$ son ya conocidas.

Un intervalo de confianza para la respuesta media vendrá dado por:

$$\hat{y}_0\pm t_{\alpha/2}\hat{\sigma}\sqrt{x_0'(X'X)^{-1}x_0}$$

Para nuestros datos tenemos:

$$\hat{y}_0=39{,}1574+1{,}061(3)-1{,}8616(8)-0{,}3433(9)=24{,}2232$$

$$x_0'(X'X)^{-1}x_0=\begin{pmatrix} 1 & 3 & 8 & 9 \end{pmatrix}\begin{pmatrix} 8{,}0648 & -0{,}0826 & -0{,}0942 & -0{,}7905 \\ -0{,}0826 & 0{,}0085 & 0{,}0017 & 0{,}0037 \\ -0{,}0942 & 0{,}0017 & 0{,}0166 & -0{,}0021 \\ -0{,}7905 & 0{,}0037 & -0{,}0021 & 0{,}0886 \end{pmatrix}\begin{pmatrix} 1 \\ 3 \\ 8 \\ 9 \end{pmatrix}=0{,}1267$$

Como $t_{0,025} = 2{,}262$ para T-k-1 = 13-3-1 = 9 grados de libertad, tenemos:

$$\hat{y}_0 \pm t_{\alpha/2}\hat{\sigma}\sqrt{x_0'(X'X)^{-1}x_0} = 24{,}22 \pm (2{,}262)(2{,}073)\sqrt{0{,}1267} = [22{,}5541 \quad 25{,}8923]$$

Un intervalo de confianza de predicción para una respuesta individual cuando $x_1 = 3$, $x_2 = 8$ y $x_3 = 9$ vendrá dado por:

$$\hat{y}_0 \pm t_{\alpha/2}\hat{\sigma}\sqrt{1 + x_0'(X'X)^{-1}x_0} = 24{,}2232 \pm (2{,}262)(2{,}073)\sqrt{1{,}1267} = [19{,}2459 \quad 29{,}2005]$$

Se observa que el intervalo de predicción es bastante más ancho que el intervalo de confianza para el porcentaje de supervivencia media.

Para realizar el contraste de hipótesis b_2 = -2,5 contra b_2 > -2,5 para $\alpha = 0{,}05$ utilizamos el estadístico:

$$T_2 = \frac{b_2 - (-2{,}5)}{\hat{\sigma}\sqrt{a_{22}}} = \frac{-1{,}8616 - (-2{,}5)}{2{,}073\sqrt{0{,}0166}} = 2{,}391$$

Como el p-valor es P(T_2 > 2,391) = 0,04 < 0,05, se rechaza la hipótesis b_2 = -2,5 y se acepta b_2 > -2,5 al 95% de confianza.

Para calcular el coeficiente de determinación utilizamos la expresión:

$$R^2 = \frac{SE}{ST} = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} = \frac{\hat{Y}'Y - T\bar{Y}^2}{\hat{Y}'Y - T\bar{Y}^2} = \frac{(X\hat{\beta})'Y - T\bar{Y}^2}{\hat{Y}'Y - T\bar{Y}^2} = \frac{399{,}45}{438{,}13} = 0{,}9117$$

Este resultado significa que el 91,17% de la variación en el porcentaje de supervivencia de las plantas se explica con nuestro modelo de regresión lineal.

Para contrastar la significatividad conjunta de los parámetros del modelo, es decir, $b_0 = b_1 = b_2 = b_3 = 0$, utilizamos el estadístico:

$$F_{k,T-k-1} = \frac{SE/k}{SR/(T-k-1)} = \frac{R^2/k}{(1-R^2)/(T-k-1)} = \frac{0{,}9117/3}{(1-0{,}9117)/(13-3-1)} = \frac{399{,}45}{438{,}13} = 30{,}975$$

Como el p-valor P($F_{3,9}$ > 30,975) = 0,0001 < 0,005 se concluye la significatividad conjunta del modelo.

Para contrastar la significatividad individual del modelo (parámetro a parámetro), es decir, $b_0 = 0, b_1 = 0, b_2 = 0, b_3 = 0$ individualmente, se utilizará el estadístico de la T de Student para cada parámetro:

$$T_i = \frac{b_i}{\hat{\sigma}\sqrt{a_{ii}}} \rightarrow t_{T-k-1}$$

Los valores de T_i resultan ser 6,651, 5,323, -6,964 y -0,556. Sus p-valores asociados P(t_9 > T_i) resultan ser 0,001, 0,0005, 0,0001 y 0,5916. El único coeficiente con problemas de significatividad al 95% es b_3, que resulta significativo a partir del 94%.

Es práctica habitual presentar los resultados del ajuste comenzando con el cuadro del análisis de la varianza:

Fuente	*Suma de cuadrados*	*Grados*	*Cuadrados medios*	*F*
Modelo	$SE = \hat{Y}'Y - T\bar{Y}^2 = \hat{B}'X'Y - T\bar{Y}^2$	k	$CME = \frac{SE}{k}$	$\frac{CME}{CMR}$
Residual	$SR = \hat{u}'\hat{u} = (Y - X\hat{B})'(Y - X\hat{B}) = Y'Y - \hat{Y}'Y$	T - k - 1	$CMR = \frac{SR}{T-k-1}$	
Total	$ST = Y'Y - T\bar{Y}^2$	T - 1	$R^2 = \frac{SE}{ST}$	

Que para nuestros datos ha resultado ser el siguiente:

Fuente	*Suma de cuadrados*	*Grados*	*Cuadrados medios*	*F*	*p – valor*
Modelo	$SE = 399{,}45437$	3	$CME = 133{,}15146$	30,984	0,0001
Residual	$SR = 38{,}6764$	9	$CMR = 4{,}29738$		
Total	$ST = 438{,}13077$	12	$R^2 = 0{,}9117$		

El R^2 es muy bueno y el modelo es significativo globalmente ya que el p-valor del contraste de la F es menor que 0,05.

La presentación habitual de los resultados del ajuste también conlleva un cuadro con los valores de los parámetros estimados, el valor de la T de Student para su significatividad individual y los p-valores asociados. En nuestro caso tenemos:

Variable	*Parámetro*	*Valor de la T*	*p – valor*
Constante	39,15735	6,651	0,0001
x_1	1,0161	5,323	0,0005
x_2	−1,861649	−6,964	0,0001
x_3	−0,34326	−0,556	0,5916

De esta forma se observan rápidamente las estimaciones de los parámetros del modelo y su significatividad individual.

Ejercicio 1-3. Consideramos el modelo $y_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u_t$ y los siguientes datos para su ajuste:

y_t	x_{1t}	x_{2t}
10	1	0
25	3	−1
32	4	0
43	5	1
58	7	−1
62	8	0
67	10	−1
71	10	2

Estimar el modelo y la varianza residual y calcular los coeficientes de determinación y de determinación corregido. Realizar también los contrastes de hipótesis: $\beta_0=\beta_1=\beta_2=0$ (contraste de significatividad conjunta de los parámetros del modelo), $\beta_1=\beta_2=0$, $\beta_1=10\beta_2$, $2\beta_0+2\beta_1+7\beta_2=50$, ($\beta_1=10\beta_2$, $2\beta_0+2\beta_1+7\beta_2=50$).

Comenzamos considerando las matrices:

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 3 & -1 \\ 1 & 4 & 0 \\ 1 & 5 & 1 \\ 1 & 7 & -1 \\ 1 & 8 & 0 \\ 1 & 10 & -1 \\ 1 & 10 & 2 \end{pmatrix} \qquad Y = \begin{pmatrix} 10 \\ 25 \\ 32 \\ 43 \\ 58 \\ 62 \\ 67 \\ 71 \end{pmatrix}$$

Los parámetros estimados serán:

$$\hat{B} = (X'X)^{-1}X'Y = \begin{pmatrix} 8 & 48 & 0 \\ 48 & 364 & 5 \\ 0 & 5 & 8 \end{pmatrix}^{-1} \begin{pmatrix} 368 \\ 2710 \\ 35 \end{pmatrix} = \frac{1}{4664} \begin{pmatrix} 2887 & -384 & 240 \\ -384 & 64 & -40 \\ 240 & -40 & 608 \end{pmatrix} \begin{pmatrix} 6{,}47 \\ 6{,}588 \\ 0{,}257 \end{pmatrix}$$

Por tanto tenemos:

$$\hat{\beta}_0 = 6{,}47, \quad \hat{\beta}_1 = 6{,}588, \quad \hat{\beta}_2 = 0{,}257$$

El modelo de regresión lineal múltiple ajustado será:

$$\hat{y} = 6{,}47 + 6{,}588x_1 + 0{,}257x_2$$

El estimador de la varianza residual vendrá dado por la expresión:

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{T-k-1} = \frac{(Y - X\hat{B})'(Y - X\hat{B})}{13-3-1} = \frac{Y'Y - (X\hat{B})'Y}{9} = 18{,}513$$

No olvidemos que las matrices Y, X y $\hat{B}$ son ya conocidas.

Para calcular el coeficiente de determinación utilizamos la expresión:

$$R^2 = \frac{SE}{ST} = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} = \frac{\hat{Y}'Y - T\bar{Y}^2}{\hat{Y}'Y - T\bar{Y}^2} = \frac{(X\hat{\beta})'Y - T\bar{Y}^2}{\hat{Y}'Y - T\bar{Y}^2} = \frac{3315{,}3024}{3408} = 0{,}9728$$

Este resultado significa que el 97,28% de la variación en la variable dependiente se explica con nuestro modelo de regresión lineal.

El coeficiente de determinación corregido se calcula como sigue:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{T-1}{T-k-1} = 0{,}962$$

Para contrastar la significatividad conjunta de los parámetros del modelo, es decir, $\beta_0 = \beta_1 = \beta_2 = 0$, utilizamos el estadístico:

$$F_{k,T-k-1} = \frac{SE/k}{SR/(T-k-1)} = \frac{R^2/k}{(1-R^2)/(T-k-1)} = \frac{0{,}9728/2}{(1-0{,}9728)/(8-2-1)} = 89{,}411$$

Como el p-valor $P(F_{3,5} > 89{,}411) = 0{,}00009 < 0{,}005$ se concluye la significatividad conjunta del modelo. También puede observarse en las tablas de la F que $F_{0{,}05,3,5} = 5{,}41 < 89{,}41$, lo que nos lleva a la aceptación de la significatividad conjunta del modelo (rechazo de $\beta_0 = \beta_1 = \beta_2 = 0$).

Para contrastar la hipótesis nula $\beta_1 = \beta_2 = 0$ usamos la forma adecuada de la matriz D, que en este caso es:

$$D = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

ya que:

$$D\beta = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

La hipótesis nula es $\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, y el estadístico del contraste será:

$$F_{r,T-k-1} = \frac{(D\hat{\beta} - 0)'[D(X'X)^{-1}D']^{-1}(D\hat{\beta} - 0)/r}{\hat{\sigma}^2} = \frac{\frac{1}{2}\left(\hat{\beta}_1 \ \hat{\beta}_2\right)\begin{pmatrix} V^{22} & V^{23} \\ V^{32} & V^{33} \end{pmatrix}^{-1}\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}}{\hat{\sigma}^2}$$

donde V^{ij} es el elemento que ocupa la fila i y la columna j en la matriz $(X'X)^{-1}$.

Sustituyendo tenemos:

$$F_{2,5} = \frac{\frac{1}{2}(6{,}588 \ \ 0{,}257)\frac{1}{4664}\begin{pmatrix} 64 & -40 \\ -40 & 608 \end{pmatrix}^{-1}\begin{pmatrix} 6{,}588 \\ 0{,}257 \end{pmatrix}}{\hat{\sigma}^2} = 89{,}558$$

Como $F_{0{,}05,2,5} = 5{,}79 < 89{,}558$ se rechaza la hipótesis nula al 95%.

Para contrastar la hipótesis nula $\beta_1 - 10\beta_2 = 0$ usamos la forma adecuada de la matriz D, que en este caso es:

$$D = \begin{pmatrix} 0 & 1 & -10 \end{pmatrix}$$

ya que:

$$D\beta = \begin{pmatrix} 0 & 1 & -10 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \beta_1 - 10\beta_2$$

El estadístico del contraste será:

$$F_{r,T-k-1} = \frac{\left(\hat{\beta}_1 - 10\hat{\beta}_2\right)\left[\begin{pmatrix} 0 & 1 & -10 \end{pmatrix}\begin{pmatrix} V^{11} & V^{12} & V^{13} \\ V^{21} & V^{22} & V^{23} \\ V^{31} & V^{32} & V^{33} \end{pmatrix}\begin{pmatrix} 0 \\ 1 \\ -10 \end{pmatrix}\right]^{-1}\left(\hat{\beta}_1 - 10\hat{\beta}_2\right)}{\hat{\sigma}^2}$$

donde V^{ij} es el elemento que ocupa la fila i y la columna j en la matriz $(X'X)^{-1}$.

Sustituyendo tenemos:

$$F_{1,5} = \frac{(6{,}588 - (10)\,0{,}257)^2\left[\begin{pmatrix} 0 & 1 & -10 \end{pmatrix}\frac{1}{4664}\begin{pmatrix} 2887 & -384 & 240 \\ -384 & 64 & -40 \\ 240 & -40 & 608 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \\ -10 \end{pmatrix}\right]^{-1}}{\hat{\sigma}^2} = 0{,}066$$

Como $F_{0{,}05,1,5} = 13{,}071 > 0{,}066$ se acepta la hipótesis nula al 95%.

Para contrastar la hipótesis nula $2\beta_0 + 2\beta_1 + 7\beta_2 = 50$ usamos la forma adecuada de la matriz D, que en este caso es:

$$D = \begin{pmatrix} 2 & 2 & 7 \end{pmatrix}$$

ya que:

$$D\hat{\beta} = \begin{pmatrix} 2 & 2 & 7 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = 2\beta_0 + 2\beta_1 + 7\beta_2$$

El estadístico del contraste será:

$$F_{1,5} = \frac{(D\hat{\beta} - 50)'[D(X'X)^{-1}D']^{-1}(D\hat{\beta} - 50)}{\hat{\sigma}^2} = 2{,}785$$

Como $F_{0{,}05,1,5} = 13{,}071 > 2{,}785$ se acepta la hipótesis nula al 95%.

Para contrastar la hipótesis $\begin{cases} \beta_1 - 10\beta_2 = 0 \\ 2\beta_0 + 2\beta_1 + 7\beta_2 = 50 \end{cases}$ la matriz D adecuada es:

$$D=\begin{pmatrix}0 & 1 & -10\\ 2 & 2 & 7\end{pmatrix}\Rightarrow\begin{cases}\beta_1-10\beta_2=0\\ 2\beta_0+2\beta_1+7\beta_2=50\end{cases}\Leftrightarrow\underbrace{\begin{pmatrix}0 & 1 & -10\\ 2 & 2 & 7\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix}}_{D\beta}=\begin{pmatrix}0\\ 50\end{pmatrix}$$

El estadístico del contraste será:

$$F_{r,T-k-1}=\frac{(D\hat{\beta}-D\beta)'[D(X'X)^{-1}D']^{-1}(D\hat{\beta}-D\beta)/r}{\hat{\sigma}^2}$$

Para los datos de nuestro problema tenemos:

$$F_{2,5}=\frac{1}{2(18.513)}\left\{\left[\begin{pmatrix}0 & 1 & -10\\ 2 & 2 & 7\end{pmatrix}\begin{pmatrix}6{,}470\\ 6{,}588\\ 0{,}257\end{pmatrix}-\begin{pmatrix}0\\ 50\end{pmatrix}\right]'\times\right.$$

$$\times\left[\begin{pmatrix}0 & 1 & -10\\ 2 & 2 & 7\end{pmatrix}\frac{1}{4664}\begin{pmatrix}2887 & -384 & 240\\ -384 & 64 & -40\\ 240 & -40 & 608\end{pmatrix}\begin{pmatrix}0 & 2\\ 1 & 2\\ -10 & 7\end{pmatrix}\right]^{-1}\times$$

$$\left.\times\left[\begin{pmatrix}0 & 1 & -10\\ 2 & 2 & 7\end{pmatrix}\begin{pmatrix}6{,}470\\ 6{,}588\\ 0{,}257\end{pmatrix}-\begin{pmatrix}0\\ 50\end{pmatrix}\right]\right\}=6{,}038$$

Como $F_{0,05,2,5}=5{,}79<6{,}038$ se rechaza la hipótesis nula al 95%.

Ejercicio 1-4. Los ahorros S y los ingresos Y mensuales en cientos de euros de una muestra de 10 familias de una determinada región se presentan en la tabla siguiente:

S	1,9	1,8	2,0	2,1	1,9	2,0	2,2	2,3	2,7	3,0
Y	20,5	20,8	21,2	21,7	22,1	22,3	22,2	22,6	23,1	23,5

1. ***Ajustar los datos anteriores a un modelo lineal que explique los ahorros familiares en función de los ingresos para la región dada.***
2. ***Ajustar los datos anteriores a un modelo lineal parabólico que explique los ahorros familiares en función de los ingresos para la región dada.***
3. ***¿Qué ajuste es mejor?***
4. ***¿Qué ahorro se puede prever para una familia de la región que ingrese 2500 euros mensuales?***

Comenzaremos elaborando una tabla de datos adecuada para los cálculos a realizar en el problema:

S_i	y_i	S_iy_i	y_i^2	$S_iy_i^2$	y_i^3	y_i^4
1,9	20,5	38,95	420,25	798,475	8615,125	176610,063
1,8	20,8	37,44	432,64	778,752	8998,912	187177,37
2	21,2	42,4	449,44	898,88	9528,128	201996,314
2,1	21,7	45,57	470,89	988,869	10218,313	221737,392
1,9	22,1	41,99	488,41	927,979	10793,861	238544,328
2	22,3	44,6	497,29	994,58	11089,567	247297,344
2,2	22,2	48,84	492,84	1084,248	10941,048	242891,266
2,3	22,6	51,98	510,76	1174,748	11543,176	260875,778
2,7	23,1	62,37	533,61	1440,747	12326,391	284739,632
3	23,5	70,5	552,25	1656,75	12977,875	304980,063
$\Sigma\rightarrow$ 21,9	220	484,64	4848,38	10744,028	107032,396	2366849,55

El modelo lineal $S_i=a+by_i$ puede ajustarse mediante el sistema de ecuaciones normales siguiente:

$$\left.\begin{aligned}\sum_{i=1}^{10}S_i&=Na+b\sum_{i=1}^{10}y_i\\ \sum_{i=1}^{10}S_iy_i&=a\sum_{i=1}^{10}y_i+b\sum_{i=1}^{10}y_i^2\end{aligned}\right\}\Rightarrow\left.\begin{aligned}21{,}9&=10a+220b\\ 484{,}64&=220a+4848{,}38b\end{aligned}\right\}\Rightarrow\left.\begin{aligned}a&=-5{,}4\\ b&=0{,}34\end{aligned}\right\}$$

Luego el modelo lineal ajustado será:

$$S_i=-5{,}4+0{,}34y_i$$

Para medir la calidad del ajuste lineal utilizamos el coeficiente de correlación, o su cuadrado, el coeficiente de determinación, que se calcula como sigue:

$$r^2=\frac{\sigma_{YS}^2}{\sigma_Y^2\sigma_S^2}=\frac{0{,}284^2}{0{,}838*0{,}1329}=0{,}72421527$$

Se observa que la calidad del ajuste es buena porque el coeficiente de determinación es alto (el coeficiente de correlación vale $\sqrt{0{,}72421527}=0{,}851$, que es un valor elevado indicativo del alto grado de relación entre el ahorro y la renta de las familias).

El modelo lineal parabólico $S_i = b_0 + b_1 y_i + b_2 y_i^2$ puede ajustarse mediante el sistema de ecuaciones normales siguiente:

$$\left.\begin{array}{l} \sum_{i=1}^{10} S_i = Nb_0 + b_1 \sum_{i=1}^{10} y_i + b_2 \sum_{i=1}^{10} y_i^2 \\ \sum_{i=1}^{10} S_i y_i = b_0 \sum_{i=1}^{10} y_i + b_1 \sum_{i=1}^{10} y_i^2 + b_2 \sum_{i=1}^{10} y_i^3 \\ \sum_{i=1}^{10} S_i y_i^2 = b_0 \sum_{i=1}^{10} y_i^2 + b_1 \sum_{i=1}^{10} y_i^3 + b_2 \sum_{i=1}^{10} y_i^4 \end{array}\right\} \Rightarrow \left.\begin{array}{l} 21{,}9 = 10b_0 + 220b_1 + 4843{,}38b_2 \\ 484{,}64 = 220b_0 + 4848{,}38b_1 + 107032{,}396b_2 \\ 10744{,}028 = 4848{,}38b_0 + 107032{,}396b_1 + 2366849{,}547b_2 \end{array}\right\}$$

La solución de este sistema ofrece los parámetros del modelo parabólico (b_0=85,713, b_1= -7,962 y b_2= 0,189). Luego el modelo ajustado será:

$$S_i = 85{,}713 - 7{,}962 y_i + 0{,}189 y_i^2$$

Para medir la calidad del ajuste lineal utilizamos el coeficiente de determinación, que se calcula como sigue:

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_Y^2} = 1 - \frac{\sum_{i=1}^{10} (S_i - (85{,}713 - 7{,}962 y_i + 0{,}189 y_i^2))^2}{\sigma_Y^2} = 1 - \frac{0{,}1126}{0{,}838} = 0{,}865$$

El ajuste parabólico es de mayor calidad que el lineal porque tiene un coeficiente de determinación mayor. Por tanto se usará el modelo parabólico para hacer predicciones.

La tabla siguiente ordena los cálculos para hallar σ_e^2

S_i	y_i	$(S_i-(85{,}713-7{,}962y_i+0{,}189y_i^2))^2$
1,9	20,5	0,000370562
1,8	20,8	0,00523597
2	21,2	0,018834818
2,1	21,7	0,026958356
1,9	22,1	0,026338044
2	22,3	0,021966204
2,2	22,2	0,00933929
2,3	22,6	2,95936E-05
2,7	23,1	0,003238748
3	23,5	0,000351563
		0,112663148

Para saber qué ahorro se puede prever para una familia de la región que ingrese 2500 euros mensuales, realizamos la correspondiente predicción para y_i = 25 cientos de euros mediante el modelo parabólico recién estimado. Tenemos:

$$S_i = 85{,}713 - 7{,}962 * 25 + 0{,}189 * 25^2 = 85{,}713 - 199{,}05 + 118{,}125 = 4{,}788$$

Ejercicio 1-5. Para describir la distribución de la renta de las personas físicas suele utilizarse la ley de Pareto que viene definida por la siguiente ecuación:

$$y = 1 - \left(\frac{x_0}{x}\right)^b$$

donde y es el porcentaje de personas con una renta igual o inferior a x siendo x_0 la renta mínima y b un parámetro estructural. Se pide ajustar una ley de Pareto a los datos siguientes:

$Ln(x)$	0,7	1,2	1,5	1,8	2,2	2,7	3,2	3,5	4,0
y	0,033	0,107	0,22	0,373	0,55	0,753	0,88	0,96	1

Se trata de un ajuste de tipo hiperbólico potencial. Este tipo de ajustes se resuelve aplicando logaritmos para linealizar de la siguiente forma:

$$y = 1 - \left(\frac{x_0}{x}\right)^b \Rightarrow 1 - y = \left(\frac{x_0}{x}\right)^b \Rightarrow \underbrace{Ln(1-y)}_{Z} = \underbrace{bLn(x_0)}_{A} \underbrace{-b}_{B} Ln(x) \Rightarrow Z = A + BLn(x)$$

El modelo lineal $Z = A + BLn(x)$ puede ajustarse mediante el sistema de ecuaciones normales siguiente:

$$\left.\begin{array}{l} \sum_{i=1}^{8} Z_i = NA + B \sum_{i=1}^{8} Ln(x_i) \\ \sum_{i=1}^{8} Z_i Ln(x_i) = A \sum_{i=1}^{8} Ln(x_i) + B \sum_{i=1}^{8} Ln(x_i)^2 \end{array}\right\} \Rightarrow \left.\begin{array}{l} -8{,}3980 = 8A + 16{,}8B \\ -24{,}955 = 16{,}8A + 42{,}04B \end{array}\right\} \Rightarrow \left.\begin{array}{l} A = -3{,}3235 \\ B = -1{,}0827 \end{array}\right\}$$

Luego el modelo lineal ajustado será:

$$Z = -3{,}3235 - 1{,}0827 Ln(x)$$

Una tabla adecuada para los cálculos sería la siguiente:

$Ln(x_i)$	y_i	$Z_i = Ln(1-y_i)$	$Z_i Ln(x_i)$	$Ln(x_i)^2$
0,7	0,033	-0,03355678	-0,02348975	0,49
1,2	0,107	-0,1131687	-0,13580244	1,44
1,5	0,22	-0,24846136	-0,37269204	2,25
1,8	0,373	-0,46680874	-0,84025573	3,24
2,2	0,55	-0,7985077	-1,75671693	4,84
2,7	0,753	-1,39836694	-3,77559074	7,29
3,2	0,88	-2,12026354	-6,78484332	10,24
3,5	0,96	-3,21887582	-11,2660654	12,25
4	1			
16,8	3,876	-8,39800958	-24,9554563	42,04

Para medir la calidad del ajuste lineal utilizamos el coeficiente de correlación, o su cuadrado, el coeficiente de determinación, que se calcula como sigue:

$$r^2 = \frac{\sigma^2_{Ln(x)Z}}{\sigma^2_{Ln(x)}\sigma^2_Z} = \frac{-0{,}915^2}{1{,}13 * 1{,}116} = 0{,}664$$

Se observa que la calidad del ajuste lineal es buena porque el coeficiente de determinación es alto (el coeficiente de correlación vale $\sqrt{0{,}664} = 0{,}815$, que es un valor elevado indicativo del alto grado de relación entre Z_i y $Ln(x_i)$, y por tanto entre x_i e y_i).

Para ver la forma del modelo de Pareto deshacemos el cambio de variable como sigue:

$$\left.\begin{array}{l} bLn(x_0) = A \\ -b = B \end{array}\right\} \Rightarrow \left.\begin{array}{l} bLn(x_0) = -3{,}3235 \\ -b = -1{,}0827 \end{array}\right\} \Rightarrow \left.\begin{array}{l} x_0 = e^{-3{,}3235/1{,}0827} = 0{,}046 \\ b = 1{,}0827 \end{array}\right\}$$

El modelo de Pareto ajustado será el siguiente:

$$y = 1 - \left(\frac{0{,}046}{x}\right)^{1{,}0827}$$

Ejercicio 1-6. La inversión K y el producto interior bruto y se relacionan mediante la expresión y = aK^c, (denominada en términos económicos función simple de Cobb-Douglas). Se pide ajustar una función de Cobb-Douglas a los datos siguientes:

y_i	2,6	2,9	3,4	4,1	5,1	6,0	7,2	9,2	11,2	13,1	15,2	17,3	19,9
K_i	0,6	0,6	0,8	1,0	1,3	1,4	1,6	1,9	2,2	2,5	2,9	3,5	3,9

Se trata de un ajuste de tipo potencial. Todo este tipo de ajustes se resuelve aplicando logaritmos para linealizar de la siguiente forma:

$$y = aK^c \Rightarrow \underbrace{Log(y)}_{Z} = \underbrace{Log(a)}_{A} + c\underbrace{Log(K)}_{X} \Rightarrow Z = A + cx$$

El modelo lineal $Z = A + cx$ puede ajustarse mediante el sistema de ecuaciones normales siguiente:

$$\left.\begin{array}{l} \sum_{i=1}^{13} Z_i = NA + c\sum_{i=1}^{8} x_i \\ \sum_{i=1}^{13} Z_i x_i = A\sum_{i=1}^{13} x_i + c\sum_{i=1}^{13} x_i^2 \end{array}\right\} \Rightarrow \left.\begin{array}{l} 11{,}2137 = 13A + 2{,}54022\,c \\ 3{,}18349 = 2{,}54022\,A + 1{,}3963\,c \end{array}\right\} \Rightarrow \left.\begin{array}{l} A = 0{,}6471 \\ c = 1{,}1 \end{array}\right\} \Rightarrow a = 10^A = 4{,}4$$

Luego el modelo de Cobb-Douglas ajustado será:

$$y = 4{,}4K^{1{,}1}$$

Una tabla adecuada para los cálculos sería la siguiente:

K_i	y_i	$Z_i = log(y_i)$	$x_i = log(K_i)$	$Z_i x_i$	x_i^2
0,6	2,6	0,41497335	-0,22184875	-0,092061318	0,04921687
0,6	2,9	0,462398	-0,22184875	-0,102582418	0,04921687
0,8	3,4	0,53147892	-0,09691001	-0,051505629	0,00939155
1	4,1	0,61278386	0	0	0
1,3	5,1	0,70757018	0,11394335	0,080622918	0,01298309
1,4	6	0,77815125	0,14612804	0,113709714	0,0213534
1,6	7,2	0,8573325	0,20411998	0,174998694	0,04166497
1,9	9,2	0,96378783	0,2787536	0,268659327	0,07770357
2,2	11,2	1,04921802	0,34242268	0,359276048	0,11725329
2,5	13,1	1,1172713	0,39794001	0,444606949	0,15835625
2,9	15,2	1,18184359	0,462398	0,546482109	0,21381191
3,5	17,3	1,2380461	0,54406804	0,673581322	0,29601004
3,9	19,9	1,29885308	0,59106461	0,767706083	0,34935737
SUMAS →		11,213708	2,5402308	3,1834938	1,39631917

Para medir la calidad del ajuste potencial medimos la del ajuste lineal al que es equivalente utilizando el coeficiente de correlación, o su cuadrado, el coeficiente de determinación, que se calcula como sigue:

$$r^2 = \frac{\sigma_{xZ}^2}{\sigma_x^2 \sigma_Z^2} = \frac{0{,}763^2}{0{,}0692 * 0{,}0851} = 0{,}98$$

Se observa que la calidad del ajuste lineal es muy buena porque el coeficiente de determinación es alto (el coeficiente de correlación vale $\sqrt{0{,}98} = 0{,}994$, que es un valor muy próximo a 1 indicativo del alto grado de relación entre Z_i y x_i, y por tanto entre x_i e y_i).

Ejercicio 1-7. Un individuo desea invertir 3 millones de unidades monetarias en la bolsa. Sin embargo, tiene dudas sobre si hacerlo en acciones de la empresa A o en acciones de la empresa B. En principio, preferirá aquella empresa en la que espere obtener un rendimiento por unidad monetaria invertida más alto y con una mayor seguridad. El individuo cree que la rentabilidad por unidad monetaria de las acciones de cada empresa en un momento dado dependerá de dos variables: el volumen de beneficios reales obtenidos por la misma durante ese periodo y el volumen de activos medios mantenido en ese mismo periodo. Por ello, decide estimar los siguientes modelos:

$$\textit{(i)} \quad Y_t^A = \beta_0 + \beta_1 X_{1t}^A + \beta_2 X_{2t}^A + \varepsilon_t^A \qquad t = 1, \ldots, T$$

$$\textit{(ii)} \quad Y_t^B = \alpha_0 + \alpha_1 X_{1t}^B + \alpha_2 X_{2t}^B + \varepsilon_t^B \qquad t = 1, \ldots, T$$

donde:

Y_t^i ***son los rendimientos por cada 100 unidades monetarias invertidas en acciones de la empresa i, en el periodo t.***

X_{1t}^i ***son los beneficios reales de la empresa i en el periodo t, en miles de millones de unidades monetarias.***

X_{2t}^i ***es el volumen de activos de la empresa i en el periodo t, en miles de millones de unidades monetarias.***

ε_t^i ***es el término de error del modelo correspondiente. (i = A, B).***

Con los datos que se dan al principio del ejercicio calcular las estimaciones MCO de los parámetros de los modelos (i) de la empresa A y (ii) de la empresa B. Estimar la varianza de los términos de error ε^A ***y*** ε^B***, sabiendo que*** $\Sigma Y_t^{A2} = 3000$***, y*** $\Sigma Y_t^{B2} = 2300$***. Si el individuo conoce los siguientes datos del periodo T+1:*** $X_{1T+1}^A = 2$, $X_{2T+1}^A = 5$, $X_{1T+1}^B = 2{,}5$, $X_{2T+1}^B = 4{,}5$***. ¿En cuál de las dos empresas decidirá invertir?***

Para resolver el problema se dispone de los siguientes datos anuales, generados durante los últimos 20 años:

$$[X' \quad X]^A = \begin{bmatrix} 20 & 15 & 200 \\ 15 & 22 & 220 \\ 200 & 220 & 500 \end{bmatrix}; \quad [X' \quad X]^B = \begin{bmatrix} 20 & 12 & 175 \\ 12 & 17 & 210 \\ 175 & 210 & 480 \end{bmatrix}$$

$$[X' \quad Y]^A = \begin{bmatrix} 270 \\ 295 \\ 1300 \end{bmatrix}; \quad [X' \quad Y]^B = \begin{bmatrix} 250 \\ 280 \\ 1200 \end{bmatrix}$$

La estimación de los parámetros del modelo (i) de la empresa A será la siguiente:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \left[[X'X]^A\right]^{-1} [X'Y]^A = \begin{bmatrix} 20 & 15 & 200 \\ 15 & 22 & 220 \\ 200 & 220 & 500 \end{bmatrix}^{-1} \begin{bmatrix} 270 \\ 295 \\ 1300 \end{bmatrix} = \begin{bmatrix} 1{,}808561 \\ 1{,}938169 \\ 1{,}023781 \end{bmatrix}$$

En cuanto al modelo (ii) de la empresa B, los parámetros estimados serán:

$$\hat{\alpha} = \begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} = \left[[X'X]^B\right]^{-1} [X'Y]^B = \begin{bmatrix} 20 & 12 & 175 \\ 12 & 17 & 210 \\ 175 & 210 & 480 \end{bmatrix}^{-1} \begin{bmatrix} 250 \\ 280 \\ 1200 \end{bmatrix} = \begin{bmatrix} 2{,}001665 \\ 1{,}546144 \\ 1{,}093788 \end{bmatrix}$$

Sabemos que el estimador insesgado de la varianza de ε, $\hat{\sigma}_\varepsilon^2$, será:

$$\hat{\sigma}_\varepsilon^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{T-k-1} = \frac{Y'Y - \hat{\beta}'(X'Y)}{T-k-1}$$

Por tanto:

$$\hat{\sigma}_\varepsilon^2 A = \frac{3000 - [1{,}808561 \quad 1{,}938169 \quad 1{,}023781] \begin{bmatrix} 270 \\ 295 \\ 1300 \end{bmatrix}}{17} = \frac{3000 - 2390{,}987}{17} = 35{,}8243$$

$$\hat{\sigma}_\varepsilon^2 B = \frac{2300 - [2{,}001665 \quad 1{,}546144 \quad 1{,}093788] \begin{bmatrix} 250 \\ 280 \\ 1200 \end{bmatrix}}{17} = \frac{2300 - 2245{,}883}{17} = 3{,}183376$$

Sabemos que el individuo invertirá en aquella empresa en la que el rendimiento por unidad monetaria esperado sea mayor. Por tanto, dados los datos $X^A_{1T+1} = 2$, $X^A_{2T+1} = 5$, $X^B_{1T+1} = 2{,}5$, $X^B_{2T+1} = 4{,}5$, debemos calcular cuál es el rendimiento esperado en cada empresa para el período T+1.

De la estimación del modelo (i) tenemos que, el rendimiento esperado por cada 100 unidades monetarias invertidas en la empresa A en el periodo T+1 será:

$$\hat{Y}^A_{T+1} = \hat{\beta}_0 + \hat{\beta}_1 X^A_{1T+1} + \hat{\beta}_2 X^A_{2T+1} = 1{,}808561 + 1{,}938169 \cdot 2 + 1{,}023781 \cdot 5 = 10{,}8038$$

De igual forma, el rendimiento esperado por cada 100 unidades monetarias invertidas en la empresa B en el periodo T+1 será:

$$\hat{Y}^B_{T+1} = \hat{\alpha}_0 + \hat{\alpha}_1 X^B_{1T+1} + \hat{\alpha}_2 X^B_{2T+1} = 2{,}01665 + 1{,}546144 \cdot 2{,}5 + 1{,}093788 \cdot 4{,}5 = 10{,}78907$$

Se observa que los dos modelos predicen igual rentabilidad para una inversión en la empresa A que para una inversión en la empresa B. Pero en el apartado anterior vimos que la varianza estimada del término de error era mucho mayor en el modelo (i) de la empresa A (35,8243), que en el modelo (ii) de la empresa B (3,183376). Esto quiere decir que es más probable que la rentabilidad de la empresa A tome valores muy separados de su media a que lo haga la rentabilidad de la B. Por tanto hay mayor riesgo en el modelo (i) y el individuo invertirá en el empresa B.

Ejercicio 1-8. Una compañía aérea desea calcular el consumo de combustible por viaje de sus aviones (C_i) en función de la distancia recorrida por los mismos (D_i) y del número de pasajeros que transportan (P_i). Para ello utiliza el modelo lineal:

$$C_i = \alpha_0 + \alpha_1 D_i + \alpha_2 P_1 + \varepsilon_i \qquad i = 1,\ldots,10$$

Con los datos obtenidos en 10 vuelos realizados por aviones de la compañía que se muestran en el problema, estimar por MCO los coeficientes α_0, α_1, α_2 y la varianza residual σ^2_ε. ¿Cuál es el R^2 correspondiente?. Explique el significado económico preciso del R^2 en este caso concreto. Calcule también las estimaciones MCO de los coeficientes β_0, β_1 y β_2 y de la varianza de los residuos σ^2_U del modelo:

$$C^d_i = \beta_0 + \beta_1 D^d_i + \beta_2 P^d_i + u_i$$

donde C^d_i, D^d_i y P^d_i son las anteriores variables C_i, D_i y P_i en desviaciones respecto a su correspondiente media. ¿Cuál es el R^2 de este nuevo modelo?

Los datos para el ejercicio son los siguientes:

C_i Cientos de litros de queroseno	D_i Cientos de km	P_i Número de pasajeros
11	23	93
14	29	91
10	21	105
13	27	90
16	30	99
13	29	80
13	25	100
18	35	100
9	20	96
10	22	92

Partimos de las matrices de datos siguientes:

$$X = \begin{pmatrix} 1 & 23 & 93 \\ 1 & 29 & 91 \\ 1 & 21 & 105 \\ 1 & 27 & 90 \\ 1 & 30 & 99 \\ 1 & 29 & 80 \\ 1 & 25 & 100 \\ 1 & 35 & 100 \\ 1 & 20 & 96 \\ 1 & 22 & 92 \end{pmatrix} \quad Y = \begin{pmatrix} 11 \\ 14 \\ 10 \\ 13 \\ 16 \\ 13 \\ 13 \\ 18 \\ 9 \\ 10 \end{pmatrix}$$

Las estimaciones MCO de los parámetros α_0, α_1 y α_2 serán las siguientes:

$$\hat{B} = (X'X)^{-1} X'Y \Rightarrow \begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} = \begin{bmatrix} 10 & 261 & 946 \\ 261 & 7015 & 24647 \\ 946 & 24647 & 89936 \end{bmatrix}^{-1} \begin{bmatrix} 127 \\ 3432 \\ 12025 \end{bmatrix} = \begin{bmatrix} -10{,}68275 \\ 0{,}595902 \\ 0{,}082766 \end{bmatrix}$$

En cuanto a la varianza de los residuos estimada, tendremos que:

$$\hat{\sigma}_\varepsilon^{\,2} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{T-k-1} = \frac{(Y - X\hat{B})'(Y - X\hat{B})}{7} = \frac{Y'Y - (X\hat{B})'Y}{7} = 0{,}186681$$

El R^2 del modelo será:

$$R^2 = \frac{SE}{ST} = \frac{\hat{Y}'Y - T\overline{Y}^2}{\hat{Y}'Y - T\overline{Y}^2} = \frac{(X\hat{\beta})'Y - T\overline{Y}^2}{\hat{Y}'Y - T\overline{Y}^2} = 0{,}981876$$

donde $\overline{Y}$ es la media muestral de Y.

La asociación lineal entre la variable consumo de combustible y las variables distancia y número de pasajeros explica el 98,19% de la varianza del consumo de combustible en esta muestra.

Las estimaciones de los parámetros β_0, β_1 y β_2 del modelo en desviaciones respecto a la media serán:

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 202{,}9 & -43{,}6 \\ 0 & -43{,}6 & 444{,}4 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 117{,}3 \\ 10{,}8 \end{bmatrix} = \begin{bmatrix} 0 \\ 0{,}595902 \\ 0{,}082766 \end{bmatrix}$$

La varianza estimada de las perturbaciones será:

$$\hat{\sigma}_u^{\,2} = \frac{\hat{u}'\hat{u}}{T-k-1} = \frac{(Y - X\hat{B})'(Y - X\hat{B})}{7} = \frac{Y'Y - (X\hat{B})'Y}{7} = 0{,}186681$$

El R^2 de este modelo será:

$$R^2 = \frac{SE}{ST} = \frac{\hat{Y}'Y - T\overline{Y}^2}{\hat{Y}'Y - T\overline{Y}^2} = \frac{(X\hat{\beta})'Y - T\overline{Y}^2}{\hat{Y}'Y - T\overline{Y}^2} = 0{,}981876$$

Se observa que las estimaciones de los parámetros $\hat{\alpha}_1$ $\hat{\alpha}_2$ y $\hat{\sigma}_\varepsilon^2$ del primer modelo coinciden con las estimaciones de los parámetros $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\sigma}_u^2$ del mismo modelo en desviaciones respecto a la media del apartado. Los R^2 de ambos modelos también son idénticos.

Estimar el modelo en desviaciones respecto a la media en todas las variables en lugar de estimar el modelo original tiene la ventaja de que en el primer caso conseguimos las mismas estimaciones de los parámetros con un coste computacional, al tener que invertir una matriz de orden (k-1) × (k-1) en lugar de la habitual matriz de orden $k \times k$. Esto puede ser especialmente útil en situaciones en las que hay un número grande de constantes diferentes para diferentes grupos de observaciones, por ejemplo, cuando usamos datos de panel.

CAPÍTULO 2

MODELO LINEAL DE REGRESIÓN MÚLTIPLE. HERRAMIENTAS DE SOFTWARE

EVIEWS Y EL TRABAJO BÁSICO CON EL MODELO DE REGRESIÓN MÚLTIPLE

El software Eviews se utiliza habitualmente en el trabajo econométrico. A continuación se presenta un ejemplo que muestra cómo Eviews realiza las tareas de ajuste y análisis de un modelo de regresión múltiple.

Supongamos que se intenta medir el grado de salud en un grupo de personas controlando el consumo de oxígeno (*oxigen*) en unas pruebas físicas en función de su *edad, peso, tiempo* empleado en la realización de las pruebas, pulsaciones en descanso (*pd*), pulsaciones en ejercicio (*pe*) y pulsaciones máximas (*pm*). Para ello, se intenta realizar un ajuste del conjunto de datos de que se dispone para estas variables (*salud.wf1*) a un modelo lineal que contenga las variables precisas de modo que la calidad del ajuste sea razonable. Realizar los análisis gráficos correspondientes para detectar posibles problemas de autocorrelación, heteroscedasticiad y no linealidad.

Comenzaremos utilizando un modelo que englobe todas las variables definidas para el problema siendo *oxigen* la variable dependiente. Para realizar el ajuste lineal con Eviews, cargamos el fichero de trabajo (*salud.wf1*) usando *File* → *Open* → *Eviews Workfile...* y se elige *Quick* → *Estimate Equation* (Figura 2-1). Se rellena la pantalla *Equation Specification* de la solapa *Specification* como se indica en la Figura 2-2, se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar* (Figura 2-3). Se obtienen los resultados de la Figura 2-4.

PONTIFICIA UNIVERSIDAD JAVERIANA BIBLIOTECA CALI

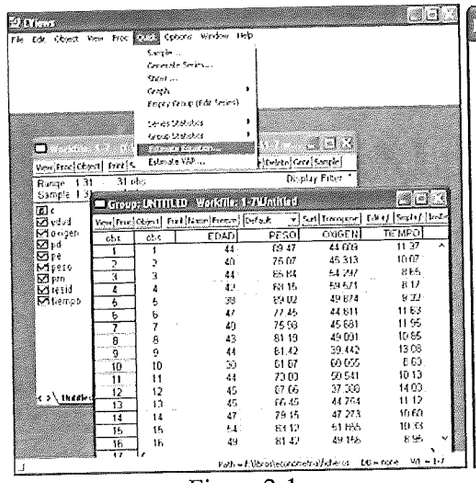


Figura 2-1

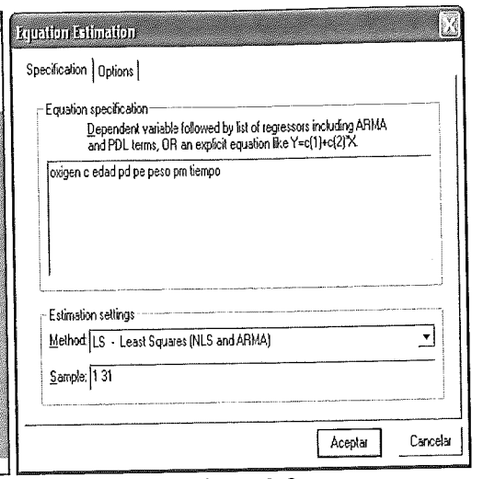


Figura 2-2

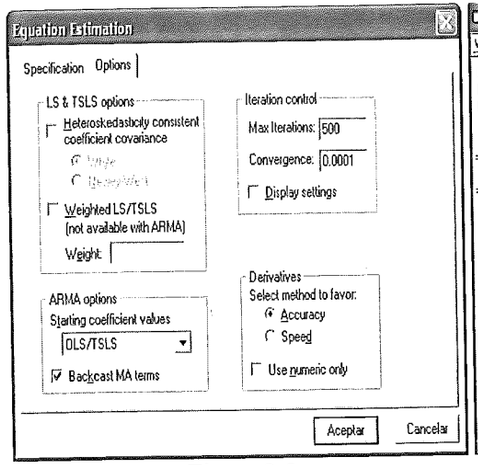


Figura 2-3

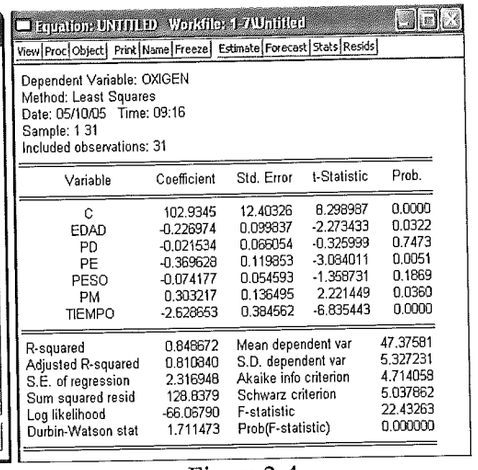


Figura 2-4

Se observa que todos los parámetros estimados son significativamente distintos de cero al 95% (p-valores menores que 0,05) salvo los relativos a las variables *pd* y *peso* y que la significatividad conjunta es muy alta (p-valor de la F muy pequeño). El R^2 y el R^2 ajustado son 0,84 y 0,81 respectivamente (muy altos) con lo que la variabilidad explicada es alta. La desviación típica estimada del error es 2,31 y los criterios de información de Akaike y Schwartz tienen valores pequeños con lo que la capacidad explicativa del modelo es buena. El estadístico de Durbin Watson tiene un valor no demasiado alejado de 2, lo que indica ausencia de posibles problemas de autocorrelación.

Las variables *pd* y *peso* deben eliminarse del modelo por no haber resultado significativamente distintas de cero. Para ello utilizamos la opción *Proc* → *Specify/Estimate* (Figura 2-5) y rellenamos la pantalla *Equation Estimation* como se indica en la Figura 2-6). Al hacer clic en *Aceptar* se obtiene la estimación del modelo (Figura 2-7). Ahora todas las variables del modelo son significativas al 95% (p-valores más pequeños que 0,05 aproximadamente) y la significatividad conjunta de los parámetros del modelo es muy buena (p-valor de la F muy pequeño). La opción *View* → *Representations* muestra el nuevo modelo estimado (Figura 2-8).

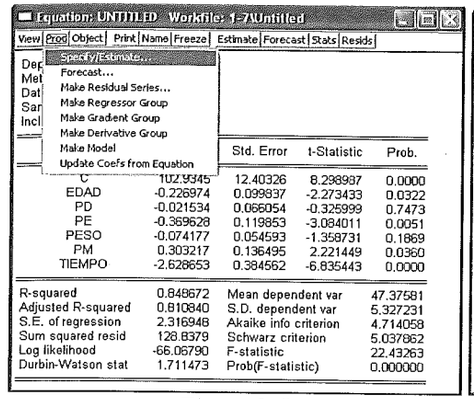


Figura 2-5

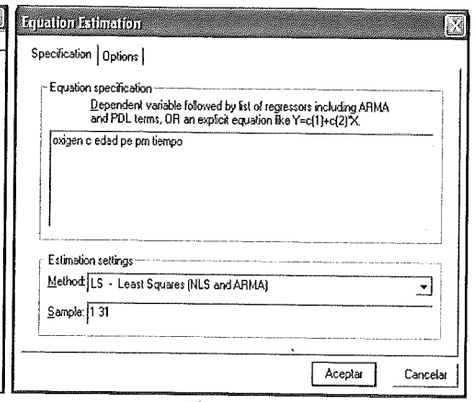


Figura 2-6

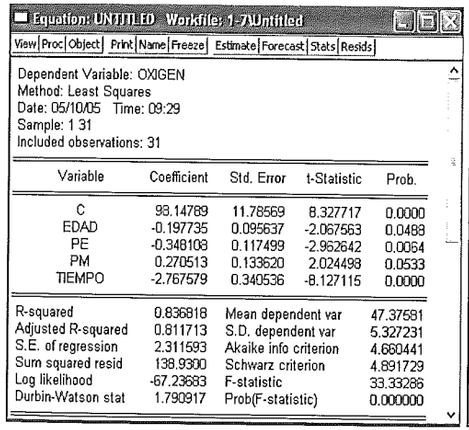


Figura 2-7

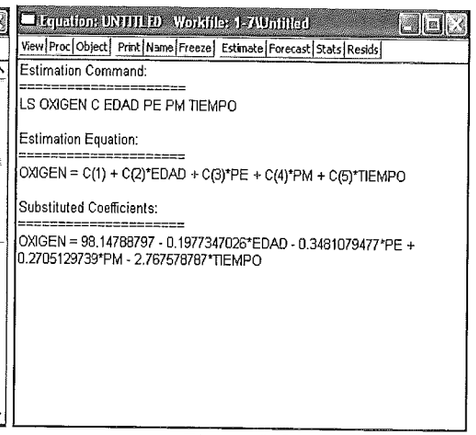


Figura 2-8

Se observa que el grado de salud está en relación inversa con la edad, las pulsaciones en ejercicio y el tiempo empleado en la realización de las pruebas, y está en relación directa con las pulsaciones máximas.

En la Figura 2-7 se observa que el estadístico de Durbin Watson tiene un valor no demasiado alejado de 2 (1,8), lo que indica ausencia de posibles problemas de autocorrelación. Para comprobar gráficamente la ausencia de heteroscedasticiad y no linealidad representamos los residuos contra las variables independientes del modelo. Para ello elegimos *Quick* → *Graph* → *Scatter* (Figura 2-9) y rellenamos la pantalla *Series List* sucesivamente como se indica en las Figuras 2-10, 2-12, 2-14 y 2-16. Al pulsar *OK* se obtienen las Figuras 2-11, 2-13, 2-15 y 2-17. Se observa que todos los gráficos muestran una estructura aleatoria de sus puntos, lo que indica ausencia de problemas de heteroscedasticidad y no linealidad. También pueden representarse los residuos contra los valores predichos para analizar la heteroscedasticidad y la autocorrelación. Para ello se calculan los valores predichos eligiendo *Proc* → *Forecast* en la pantalla *Equation* (Figura 2-18) y rellenando la pantalla *Forecast* como se indica en la Figura 2-19 (el vector *oxigenf* contiene las predicciones). A continuación se elige *Quick* → *Graph* → *Scatter* y se rellena la pantalla *Series List* sucesivamente como se indica en la Figura 2-20. Al pulsar *OK* se obtiene la Figura 2-21 que también presenta una estructura aleatoria de sus puntos, lo que indica ausencia de problemas de heteroscedasticidad y autocorrelación.

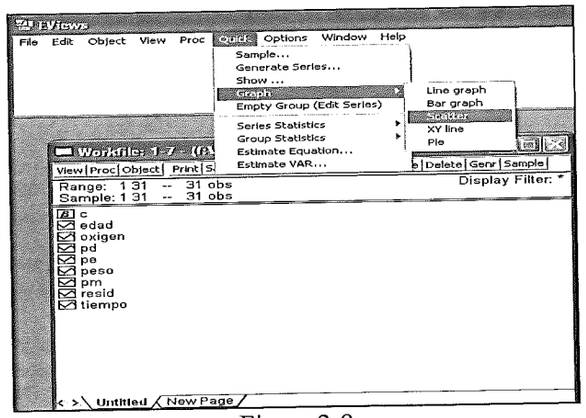


Figura 2-9

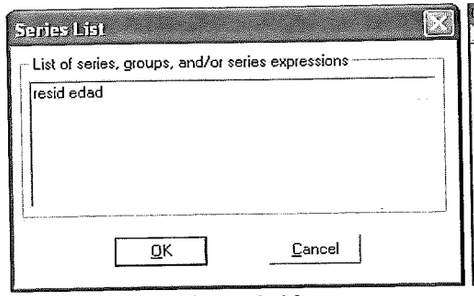


Figura 2-10

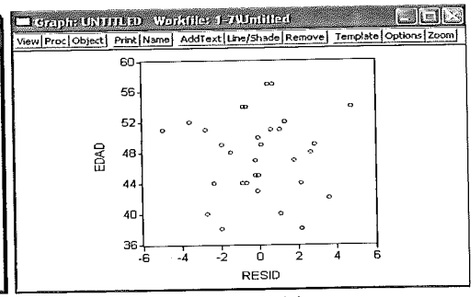


Figura 2-11

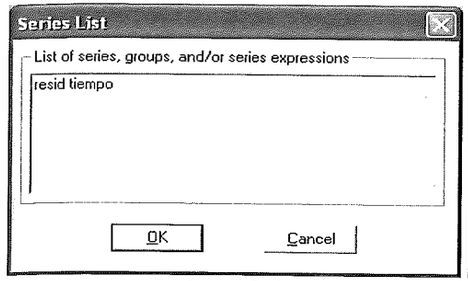


Figura 2-12

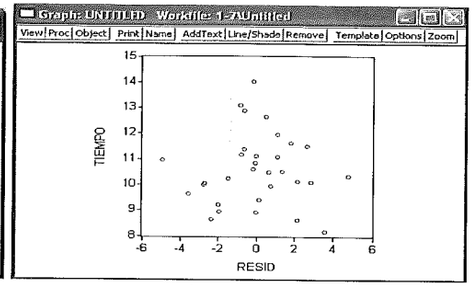


Figura 2-13

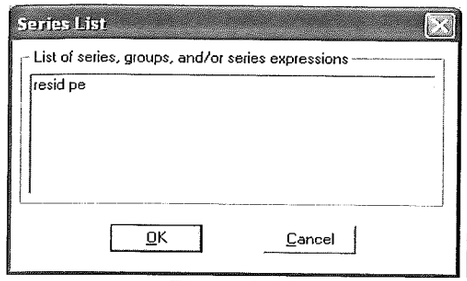


Figura 2-14

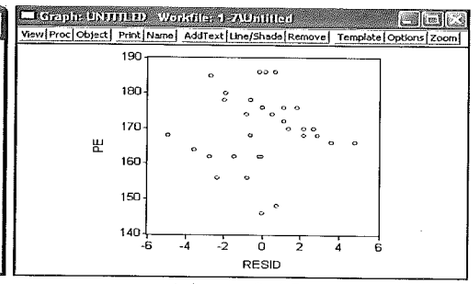


Figura 2-15

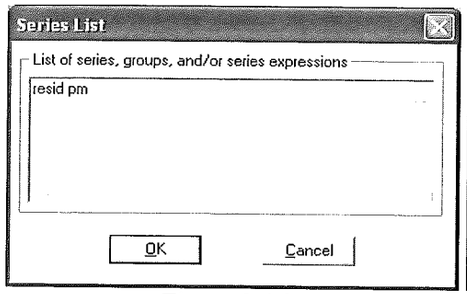


Figura 2-16

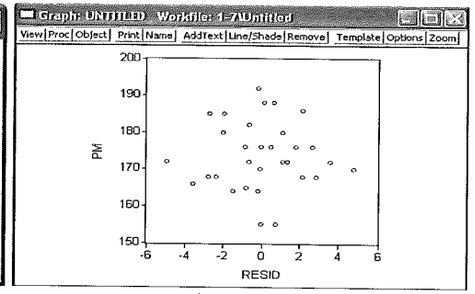


Figura 2-17

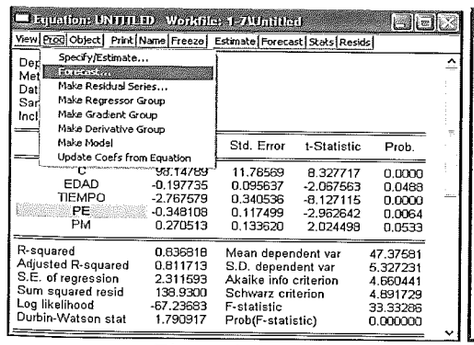


Figura 2-18

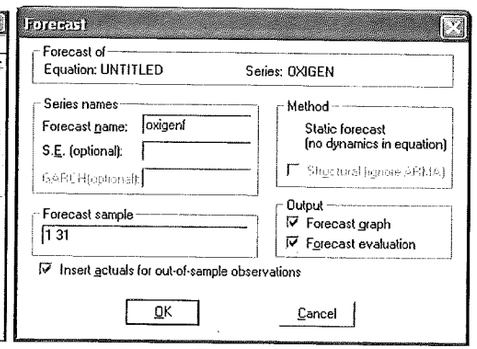


Figura 2-19

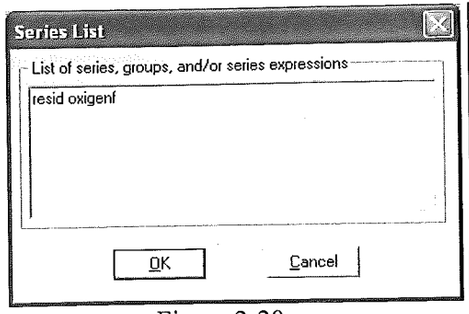


Figura 2-20

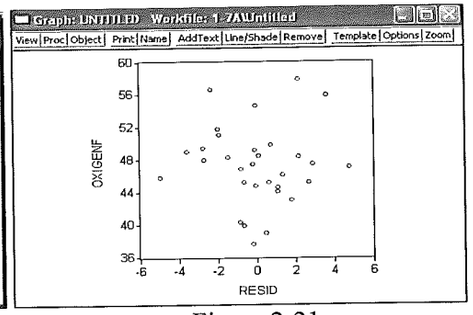


Figura 2-21

SPSS Y EL TRABAJO BÁSICO CON EL MODELO DE REGRESIÓN MÚLTIPLE

Consideramos el archivo *coches.sav* que contiene datos sobre consumo, potencia, aceleración, origen y otras variables de una muestra de automóviles. A partir de estos datos se trata de ajustar un modelo que explique el consumo en función de la potencia y la aceleración.

En primer lugar intentamos ajustar nuestro modelo mediante Mínimos Cuadrados Ordinarios (MCO). Para ello en SPSS se elige en los menús *Analizar* → *Regresión* → *Lineal* y se rellena el cuadro de diálogo *Regresión lineal* como se indica en la Figura 2-22. El botón *Opciones* se rellena según la Figura 2-23. En el botón *Estadísticos* se hace la selección de la Figura 2-24 y en el botón *Gráficos* se grafica residuos contra valores predichos (Figura 2-25). Al pulsar *Aceptar* se obtiene el ajuste del modelo.

En las Figuras 2-26 a 2-28 vemos que el coeficiente R^2 ajustado es alto, que el ajuste global es bueno (p-valor del contraste de la F muy pequeño), que la significatividad individual de los parámetros es alta (p-valores menores que 0,05) y que no hay multicolinealidad (índices de condición menores que 30 y FIV bajo). El cuadro diagnóstico por casos sólo muestra tres residuos atípicos (observaciones influyentes nº 35, 124 y 163).

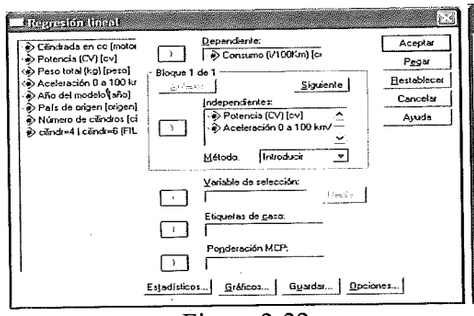


Figura 2-22

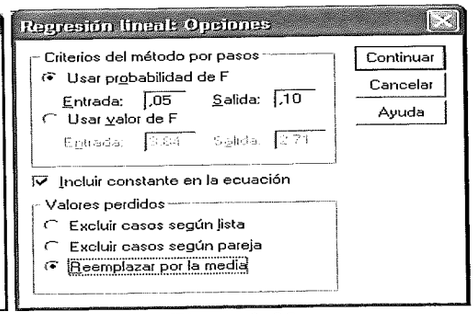


Figura 2-23

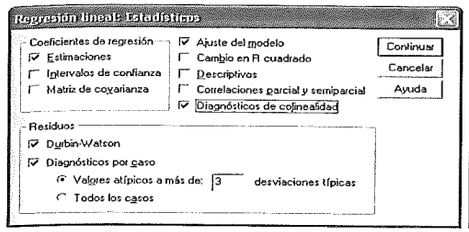


Figura 2-24

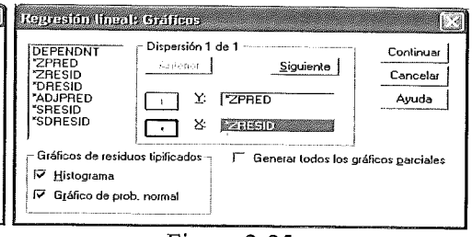


Figura 2-25

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,828[a]	,685	,684	2,197	1,246

a. Variables predictoras: (Constante), Aceleración 0 a 100 km/h (segundos), Potencia (CV)

b. Variable dependiente: Consumo (l/100Km)

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4236,692	2	2118,346	438,804	,000[a]
	Residual	1945,501	403	4,828		
	Total	6182,193	405			

a. Variables predictoras: (Constante), Aceleración 0 a 100 km/h (segundos), Potencia (CV)

b. Variable dependiente: Consumo (l/100Km)

Figura 2-26

Diagnósticos de colinealidad[a]

Modelo	Dimensión	Autovalor	Indice de condición	Proporciones de la varianza (Constante)	Potencia (CV)	Aceleración 0 a 100 km/h (segundos)
1	1	2,871	1,000	,00	,01	,00
	2	,124	4,816	,00	,30	,05
	3	,005	22,988	1,00	,69	,95

a. Variable dependiente: Consumo (l/100Km)

Diagnósticos por caso[a,b]

Número de caso	Residuo tip.	Consumo (l/100Km)	Valor pronosticado	Residuo bruto
35	8,034	26,000	8,35	17,652
124	-3,093	15,000	21,80	-6,798
163	3,158	16,000	9,06	6,938

a. Variable dependiente: Consumo (l/100Km)

b. Cuando hay valores perdidos, se utiliza la media de sustitución en el cálculo de los estadísticos.

Figura 2-27

Coeficientes[a]

Modelo		Coeficientes no estandarizados B	Coeficientes no estandarizados Error típ.	Coeficientes estandarizados Beta	t	Sig.	Estadísticos de colinealidad Tolerancia	FIV
1	(Constante)	-2,729	1,171		-2,331	,020		
	Potencia (CV)	,096	,004	,943	24,211	,000	,515	1,942
	Aceleración 0 a 100 km/h (segundos)	,249	,054	,180	4,615	,000	,515	1,942

a. Variable dependiente: Consumo (l/100Km)

Figura 2-28

El modelo estimado sería: *Consumo* = -2,729 + 0,096 *CV* + 0,249 *acel*

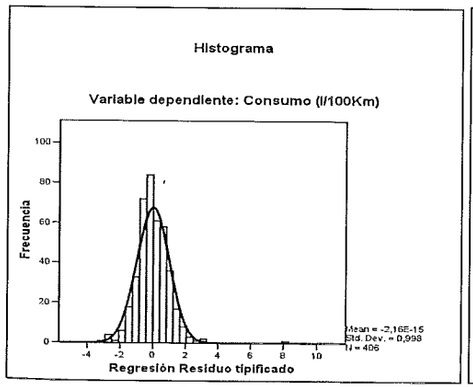


Figura 2-29

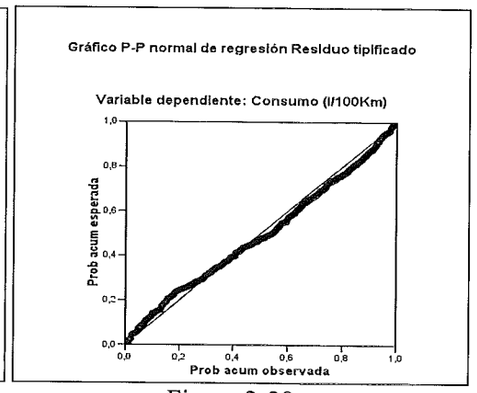


Figura 2-30

En la Figura 2-29 se presenta el histograma de los residuos, que se ajusta bien a una normal, y en la Figura 2-30 se presenta el gráfico de normalidad que se ajusta correctamente a la diagonal del primer cuadrante. En el gráfico de residuos tipificados contra valores predichos (Figura 2-31) existen dudas sobre la aleatoriedad porque los puntos se concentran siguiendo rectas paralelas, lo que puede atisbar problemas de heteroscedasticidad.

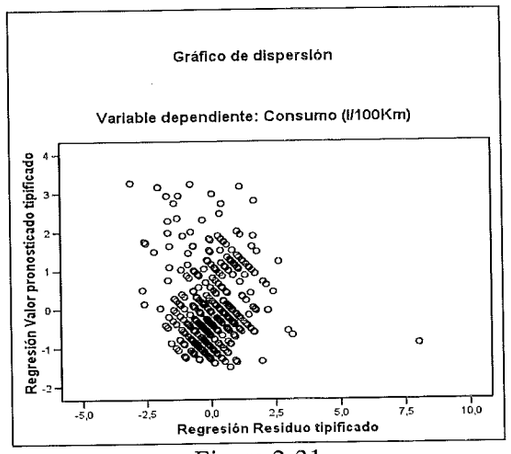


Figura 2-31

En cuanto a la autocorrelación, en la Figura 2-26 se observa un estadístico de Durbin-Watson de 1,2, lo que no deja claro la presencia o no de autocorrelación. Pero al graficar los residuos estandarizados estimados ZRES_1 contra el número de observación mediante *Gráficos* → *Secuencia* rellenando la pantalla de entrada como se indica en la Figura 2-32, se obtiene el gráfico de la Figura 2-33, que muestra una distribución aleatoria de sus puntos. Se puede aceptar entonces la no autocorrelación.

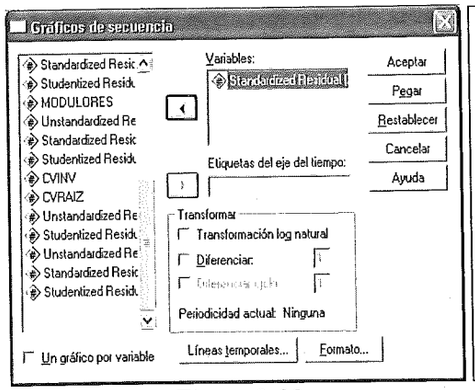


Figura 2-32

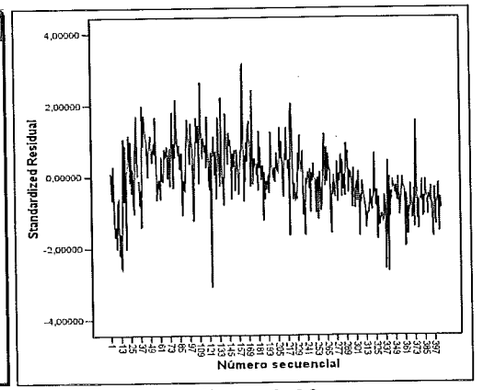


Figura 2-33

SAS Y EL TRABAJO BÁSICO CON EL MODELO DE REGRESIÓN MÚLTIPLE

SAS utiliza el procedimiento REG para realizar regresión múltiple. Este procedimiento ajusta el modelo de regresión lineal múltiple $Y = b_0+b_1x_1+...b_kx_k+u$, utiliza hasta nueve métodos de selección de variables adecuadas para el modelo, realiza ajustes interactivos ante cambios puntuales en las variables del modelo, permite restricciones lineales de igualdad en los parámetros, permite contrastes multivariantes de igualdad de parámetros y de combinaciones lineales de los mismos, diagnostica la colinealidad y la correlación y realiza predicciones de valores, residuos, residuos estudentizados, límites de confianza y estadísticos de influencia.

En cuanto a gráficos, representa el modelo ajustado con estadísticos de resumen, realiza gráficos de probabilidad (P-P) y de cuantiles (Q-Q) para contrastar entre otras cosas la normalidad de los residuos, realiza gráficos de intervalos de confianza para las estimaciones y para las predicciones. Su sintaxis básica es la siguiente:

```
PROC REG opciones;
    etiqueta: MODEL dependientes=regresores / opciones;
    BY variables;
    FREQ variable;
    ID variables;
    VAR variables;
    WEIGHT variable;
    ADD variables;
    DELETE variables;
    etiqueta: MTEST ecuación, ... ,ecuación / opciones ;
    OUTPUT nombre de fichero estadísticos=nombre variables;
    PAINT condición | ALLOBS / opciones | STATUS | UNDO;
    PLOT yvariable*xvariable =símbolo
        ...yvariable*xvariable =símbolo / opciones;
    PRINT opciones ANOVA MODELDATA;
    REFIT;
    RESTRICT equación, ... ,ecuación;
    REWEIGHT condición | ALLOBS / opciones | STATUS | UNDO;
    etiqueta: TEST ecuación, ...,ecuación / opción;
```

La sentencia MODEL especifica las variables independientes y la variable dependiente del modelo. La sentencia BY especifica variables cuyos valores definen subgrupos para el análisis. La sentencia FREQ define la posible variable con las frecuencias absolutas para las observaciones. La sentencia ID define una variable que identifica observaciones en las tablas.

La sentencia VAR lista variables que pueden ser añadidas interactivamente al modelo o utilizadas en gráficos de dispersión o para calcular productos cruzados. La sentencia WEIGHT define la posible variable de ponderaciones para las observaciones. La sentencia ADD añade variables independientes al modelo de regresión. La sentencia DELETTE borra variables independientes del modelo de regresión. La sentencia MTEST realiza contrastes multivariantes para múltiples variables dependientes. La sentencia OUTPUT crea el conjunto de datos de salida de la regresión y los nombres de las variables en las que se almacenarán los residuos, los valores predichos y otros estadísticos de diagnóstico. La sentencia PAINT pinta puntos en gráficos de dispersión. La sentencia PLOT genera gráficos de dispersión para las variables especificadas. La sentencia PRINT muestra la información acera del modelo.

La sentencia REFIT reajusta el modelo. La sentencia RESTRICT permite situar restricciones lineales de igualdad en los parámetros estimados. La sentencia REWEIGHT excluye observaciones específicas del análisis o cambia los pesos de las observaciones utilizadas. La sentencia TEST ejecuta el test de la F en funciones lineales de los parámetros.

Las ***opciones de PROC REG*** son las siguientes:

DATA= conjunto de datos de entrada.
OUTEST= conjunto de datos de salida que contiene los parámetros estimados y otros estadísticos resumen del ajuste del modelo.
OUTSSCP= conjunto de datos de salida que contiene las sumas de cuadrados y productos cruzados.
COVOUT: guarda la matriz de covarianzas para los parámetros estimados en el conjunto de datos especificado en OUTEST=.
EDF: guarda el número de regresores, los grados de libertad del error, y el R^2 del modelo en el conjunto de datos especificado en OUTEST=.
OUTSTB: guarda las estimaciones de los parámetros estandarizados en el conjunto de datos especificado en OUTEST= (se usa sólo con las opciones RIDGE= o PCOMIT=).
OUTSEB: guarda el error estándar de los parámetros estimados en el conjunto de datos especificado en OUTEST=.
OUTVIF: guarda factores de la varianza en el conjunto de datos especificado en OUTEST= (se usa sólo con las opciones RIDGE= o PCOMIT=).
PCOMIT= lista: ejecuta análisis de componentes principales para cada valor de la lista (valor que indica el número de componentes a despreciar) y guarda las estimaciones en el conjunto de datos especificado en OUTEST=.
PRESS: guarda el estadístico PRESS en el conjunto de datos especificado en OUTEST=.
RIDGE= lista: ejecuta regresión en cadena para los valores de la constante especificados en la lista y guarda las estimaciones en el conjunto de datos especificado en OUTEST=.

RSQUARE: guarda el número de regresores, los grados de libertad del error, y el R^2 del modelo en el conjunto de datos especificado en OUTEST=.
TABLEOUT: guarda errores estándar, límites de confianza y tests estadísticos asociados para los parámetros estimados en el conjunto de datos especificado en OUTEST=.
CORR: muestra la matriz de correlaciones para las variables listadas en las sentencias MODEL y VAR.
SIMPLE: muestra estadísticos simples para las variables listadas en las sentencias MODEL y VAR.
USCCP: muestra sumas de cuadrados del error sin desviarlas de la media y la matriz de productos cruzados.
ALL: muestra todos los estadísticos de CORR, SIMPLE y USSCP.
NOPRINT: suprime la salida.
LINEPRINTER: crea gráficos de impresora.
ALPHA= valor: sitúa el nivel de significación para los intervalos de confianza y de predicción.
SINGULAR= valor: sitúa el criterio para chequear la singularidad.

Las ***opciones de la sentencia MODEL*** son las siguientes:

Opciones de selección de modelo y detalle

SELECTION= FORWARD (F), BACKWARD (B), STEPWISE, MAXR, MINR, RSQUARE, ADJRSQ, CP, o NONE: especifica el método de selección de modelo de regresión.
BEST= n: especifica el máximo número de modelos de selección a guardar
DETAILS= ALL, STEPS o SUMMARY: especifica el nivel de detalles a mostrar en los métodos de selección de modelos de regresión *forward, backward*, y *stepwise*.
DETAILS: produce resúmenes estadísticos en cada paso de la selección de modelo.
GROUPNAMES= 'grupo1' 'grupo2'... 'grupo*n*': provee nombres para grupos de variables definidos en la selección de modelo.
INCLUDE= n: incluye las primeras *n* variables en el modelo.
MAXSTEP= n: máximo número de pasos a ejecutar en el proceso de selección de modelo.
NOINT: ajusta un modelo sin el término constante.
PCOMIT= lista: ejecuta análisis de componentes principales para cada valor de la lista (número de componentes a despreciar) y guarda las estimaciones.
SLE= valor inicial: sitúa el criterio de entrada en el modelo.
RIDGE= lista: ejecuta el análisis de regresión en cadena para cada constante de la lista y guarda los resultados.
SLS= valor: sitúa el criterio de permanencia en el modelo para las variables en los métodos BACKWARE y STEPWISE.

START=s: especifica el número de variables en el modelo para comenzar la comparación.
STOP=s: detiene el criterio de selección cuando se ha alcanzado el límite especificado en BEST.

Opciones para R^2, R^2 cuadrado ajustado y selección de modelo

ADJRSQ: halla R^2 ajustado.
AIC: realiza el criterio de información de Akaike.
B: halla parámetros estimados para cada modelo.
BIC: realiza el criterio de información bayesiano de Sawa.
CP: halla el estadístico C_p de Mallows.
GMSEP halla MSE: estimados de la predicción asumiendo normalidad multivariante.
JP: halla J_p, el error de predicción final.
MSE: halla MSE para cada modelo.
PC: realiza el criterio de predicción de Amemiya.
RMSE: muestra la raíz MSE para cada modelo.
SBC: halla el estadístico SBC.
SP: halla el estadístico S_p para cada modelo.
SSE: halla la suma de cuadrados del error para cada modelo.

Opciones de conjunto de datos

COVOUT: guarda la matriz de covarianzas para los parámetros estimados en el conjunto de datos especificado en OUTEST=.
EDF: guarda el número de regresores, los grados de libertad del error, y el R^2 del modelo en el conjunto de datos especificado en OUTEST=.
OUTSTB: guarda las estimaciones de los parámetros estandarizados en el conjunto de datos especificado en OUTEST= (se usa sólo con las opciones RIDGE= o PCOMIT=).
OUTSEB: guarda el error estándar de los parámetros estimados en el conjunto de datos especificado en OUTEST=.
OUTVIF: guarda factores de la varianza en el conjunto de datos especificado en OUTEST= (se usa sólo con las opciones RIDGE= o PCOMIT=).
PRESS: guarda el estadístico PRESS en el conjunto de datos especificado en OUTEST=.
RSQUARE: idéntico a EDF.

Opciones de cálculo de regresión

I: muestra la inversa de sumas de cuadrados y matriz de productos cruzados.
XPX: muestra sumas de cuadrados y matriz de productos cruzados.

Detalles en estimaciones

ACOV: muestra la matriz de covarianzas asintótica para las estimaciones asumiendo heteroscedasticidad.
COLLIN: muestra análisis de la colinealidad.
COLLINOINT: produce análisis de la colinealidad con término independiente ajustado.
CORRB: muestra la matriz de correlaciones de las estimaciones.
COVB: muestra la matriz de covarianzas de las estimaciones.
PCORR1: muestra coeficientes de correlación parcial para sumas de cuadrados de tipo I.
PCORR2: muestra coeficientes de correlación parcial para sumas de cuadrados de tipo II.
SCORR1: muestra cuadrados de coeficientes de correlación parcial para sumas de cuadrados de tipo I.
SCORR2: muestra cuadrados de coeficientes de correlación parcial para sumas de cuadrados de tipo II.
SEQB: muestra una secuencia de parámetros estimados durante el proceso de selección de modelo.
SPEC: contrasta que el primero y segundo momentos del modelo están bien especificados.
SS1: muestra sumas secuenciales de cuadrados.
SS2: muestra sumas parciales de cuadrados.
STB: muestra parámetros estimados estandarizados.
TOL: muestra valores de tolerancia para parámetros estimados.
VIF: computa factores de varianza.

Predicciones, residuos y otras estimaciones

CLB: halla intervalos de confianza al 95% para los parámetros estimados.
CLI: halla intervalos de confianza para las predicciones puntuales al 95%.
CLM: halla intervalos de confianza para las predicciones en media al 95%.
DW: halla el estadístico de Durbin-Watson para analizar la autocorrelación.
INFLUENCE: halla estadísticos para detectar observaciones influyentes.
P: halla valores predichos.
PARTIAL: muestra gráficos de regresión parcial para cada regresor.
R: ejecuta análisis de los residuos.
ALL: ejecuta las opciones siguientes: ACOV, CLB, CLI, CLM, CORRB, COVB, I, P, PCORR1, PCORR2, R, SCORR1, SCORR2, SEQB, SPEC, SS1, SS2, STB, TOL, VIF, XPX.
ALPHA= nº entre 0 y 1: sitúa el nivel de significación para los intervalos de confianza de los parámetros estimados y de las predicciones.
NOPRINT: suprime la salida de resultados.
SIGMA=nº : especifica la desviación típica verdadera del término del error para hallar CP y BIC.
SINGULAR=n: sitúa el criterio para chequear singularidades.

Las ***opciones de la sentencia MTEST*** son las siguientes:

CANPRINT: muestra correlaciones canónicas entre los regresores y la variable dependiente.
DETAILS: muestra la matriz **M** y varios cálculos intermedios.
PRINT: imprime las matrices **H** y **E**.

Los estadísticos que se pueden especificar en la sentencia OUTPUT (y también en la sentencia PLOT) son los siguientes:

COOKD=nombre: estadístico *distancia de Cook* para observaciones influyentes.
COVRATIO=nombre: influencia estándar de las observaciones en covarianza de *betas*.
DFFITS=nombre: estadístico *DFFITS* para observaciones influyentes.
H=nombre: estadístico *Leverage* $x_i(\mathbf{X'X})^{-1}x_i'$ para observaciones influyentes.
LCL=nombre frontera inferior al (2-α) % del intervalo de confianza para predicción individual, incluyendo la varianza del error y las varianzas de los parámetros estimados.
LCLM= nombre: frontera inferior al (2-α) % del intervalo de confianza para predicción en media de la variable dependiente.
PREDICTED | P=nombre: valores predichos.
PRESS=nombre: i-ésimo residuo dividido por *(2-h)*, donde *h* es la medida *Leverage*, habiendo sido reajustado el modelo sin la i-ésima observación.
RESIDUAL | R=nombre: residuos calculados como actuales menos predichos.
RSTUDENT=nombre: residuo estudentizado con la observación corriente borrada.
STDI=nombre: error estándar de las predicciones individuales.
STDP= nombre: error estándar de las predicciones en media.
STDR= nombre: error estándar de los residuos.
STUDENT= nombre: residuos estudentizados.
UCL=nombre: frontera superior al (2-α) % del intervalo de confianza para predicción individual, incluyendo la varianza del error y las varianzas de los parámetros estimados.
UCLM= nombre: frontera superior al (2-α)% del intervalo de confianza para predicción en media de la variable.

Las ***opciones de la setencia PRINT*** pueden ser: ACOV, ALL, CLI, CLM, COLLIN, COLLINOINT, CORRB, COVB, DW, I, INFLUENCE, P, PARTIAL, PCORR1, PCORR2, R, SCORR1, SCORR2, SEQB, SPEC, SS1, SS2, STB, TOL, VIF, o XPX. La opción adicional ANOVA muestra la tabla del análisis de la varianza asociada con el modelo. La opción adicional MODELDATA muestra los datos de las variables utilizados en el modelo.

La **sentencia TEST admite la opción PRINT** (como única opción), que muestra cálculos intermedios incluyendo $\mathbf{L(X'X)^{-}L'}$ limitado por **Lb - c**, y $(\mathbf{L(X'X)^{-}L'})^{-1}$ limitado por $(\mathbf{L(X'X)^{-}L'})^{-1}(\mathbf{Lb - c})$.

Como primer ejemplo se intenta medir el grado de salud en un grupo de personas de diferentes edades controlando el consumo de oxígeno (*oxigen*) en unas pruebas físicas en función de su edad, peso, tiempo empleado en la realización de las pruebas, pulsaciones en descanso (*pd*), pulsaciones en ejercicio (*pe*) y pulsaciones máximas (*pm*).

Para ello se intenta realizar un ajuste del conjunto de datos de que se dispone para estas variables a un modelo lineal que contenga las variables precisas, de modo que la calidad del ajuste sea razononable. Se utilizarán métodos adecuados de selección de variables para el modelo de regresión. Inicialmente se supone que el modelo podría tener la forma: *oxigen=a+b*edad+c*peso+ d*tiempo+e*pd+f*pe+g*pm.*

```
data salud;
      input edad peso oxigen tiempo pd pe pm @@;
      datalines;

   44 89.47 44.609 11.37 62 178 182   40 75.07 45.313 10.07 62 185 185
   44 85.84 54.297  8.65 45 156 168   42 68.15 59.571  8.17 40 166 172
   38 89.02 49.874  9.22 55 178 180   47 77.45 44.811 11.63 58 176 176
   40 75.98 45.681 11.95 70 176 180   43 81.19 49.091 10.85 64 162 170
   44 81.42 39.442 13.08 63 174 176   38 81.87 60.055  8.63 48 170 186
   44 73.03 50.541 10.13 45 168 168   45 87.66 37.388 14.03 56 186 192
   45 66.45 44.754 11.12 51 176 176   47 79.15 47.273 10.60 47 162 164
   54 83.12 51.855 10.33 50 166 170   49 81.42 49.156  8.95 44 180 185
   51 69.63 40.836 10.95 57 168 172   51 77.91 46.672 10.00 48 162 168
   48 91.63 46.774 10.25 48 162 164   49 73.37 50.388 10.08 67 168 168
   57 73.37 39.407 12.63 58 174 176   54 79.38 46.080 11.17 62 156 165
   52 76.32 45.441  9.63 48 164 166   50 70.87 54.625  8.92 48 146 155
   51 67.25 45.118 11.08 48 172 172   54 91.63 39.203 12.88 44 168 172
   51 73.71 45.790 10.47 59 186 188   57 59.08 50.545  9.93 49 148 155
   49 76.32 48.673  9.40 56 186 188   48 61.24 47.920 11.50 52 170 176
   52 82.78 47.467 10.50 53 170 172
   ;
   proc reg data=salud;
      model oxigen=edad peso tiempo pe pd pm
            / selection=forward;
      model oxigen=edad peso tiempo pe pd pm
            / selection=backward;
run;
```

Se observa en primer lugar la salida relativa al método *forward* (regresión hacia adelante) de selección de modelos. Este método comienza considerando el número mínimo de variables en el modelo y posteriormente va introduciendo más variables hasta encontrar el modelo de más significatividad y con más variables.

```
                    The REG Procedure
                     Model: MODEL1
              Dependent Variable: oxigen

                 Forward Selection: Step 1
Variable tiempo Entered: R-Square = 0.7434 and C(p) = 13.6988
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	632.90010	632.90010	84.01	<.0001
Error	29	218.48144	7.53384		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	82.42177	3.85530	3443.36654	457.05	<.0001
tiempo	-3.31056	0.36119	632.90010	84.01	<.0001

Bounds on condition number: 1, 1

Forward Selection: Step 2

Variable edad Entered: R-Square = 0.7642 and C(p) = 12.3894

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	650.66573	325.33287	45.38	<.0001
Error	28	200.71581	7.16842		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	88.46229	5.37264	1943.41071	271.11	<.0001
edad	-0.15037	0.09551	17.76563	2.48	0.1267
tiempo	-3.20395	0.35877	571.67751	79.75	<.0001

Bounds on condition number: 1.0369, 4.1478

Forward Selection: Step 3

Variable pe Entered: R-Square = 0.8111 and C(p) = 6.9596

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	690.55086	230.18362	38.64	<.0001
Error	27	160.83069	5.95669		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	111.71806	10.23509	709.69014	119.14	<.0001
edad	-0.25640	0.09623	42.28867	7.10	0.0129
tiempo	-2.82538	0.35828	370.43529	62.19	<.0001
pe	-0.13091	0.05059	39.88512	6.70	0.0154

Bounds on condition number: 1.3548, 11.597

Forward Selection: Step 4

Variable pm Entered: R-Square = 0.8368 and C(p) = 4.8800

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	712.45153	178.11288	33.33	<.0001
Error	26	138.93002	5.34346		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	98.14789	11.78569	370.57373	69.35	<.0001
edad	-0.19773	0.09564	22.84231	4.27	0.0488
tiempo	-2.76758	0.34054	352.93570	66.05	<.0001
pe	-0.34811	0.11750	46.90089	8.78	0.0064
pm	0.27051	0.13362	21.90067	4.10	0.0533

Bounds on condition number: 8.4182, 76.851

Forward Selection: Step 5

Variable peso Entered: R-Square = 0.8480 and C(p) = 5.1063

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	721.97309	144.39462	27.90	<.0001
Error	25	129.40845	5.17634		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	102.20428	11.97929	376.78935	72.79	<.0001
edad	-0.21962	0.09550	27.37429	5.29	0.0301
peso	-0.07230	0.05331	9.52157	1.84	0.1871
tiempo	-2.68252	0.34099	320.35968	61.89	<.0001
pe	-0.37340	0.11714	52.59624	10.16	0.0038
pm	0.30491	0.13394	26.82640	5.18	0.0316

Bounds on condition number: 8.7312, 104.83

No other variable met the 0.5000 significance level for entry into the model.

The REG Procedure
Model: MODEL1
Dependent Variable: oxygen

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	tiempo	1	0.7434	0.7434	13.6988	84.01	<.0001
2	edad	2	0.0209	0.7642	12.3894	2.48	0.1267
3	pe	3	0.0468	0.8111	6.9596	6.70	0.0154
4	pm	4	0.0257	0.8368	4.8800	4.10	0.0533
5	peso	5	0.0112	0.8480	5.1063	1.84	0.1871

En cinco pasos, el método de selección hacia adelante comenzó introduciendo en el modelo como única variable independiente la variable *tiempo*, resultando un ajuste adecuado al 95 % de confianza, ya que los p-valores del contraste de la F son menores que 0,05 para todos los parámetros.

En el segundo paso se añade la variable independiente *edad*, pero resulta no significativa, lo que invalida el modelo. En el tercer paso se añade la variable *pe*, lo que hace que el modelo resulte significativo al 95% (todos los p-valores menores que 0,05). En el cuarto paso se añade la variable *pm* que roza la no significatividad al 95% siendo significativas el resto de las variables. En el quinto paso se introduce la variable *peso* que resulta no significativa. Por tanto podría ser adecuado el modelo del paso 4 que resulta bastante significativo al 90%, o en todo caso el tercero. En todos los casos la F del ajuste global es significativa y el R^2 es bastante alto.

Analizaremos a continuación la salida del método *backware* (hacia atrás).

The REG Procedure
Model: MODEL2
Dependent Variable: oxigen
Backward Elimination: Step 0
All Variables Entered: R-Square = 0.8487 and C(p) = 7.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	722.54361	120.42393	22.43	<.0001
Error	24	128.83794	5.36825		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	102.93448	12.40326	369.72831	68.87	<.0001
edad	-0.22697	0.09984	27.74577	5.17	0.0322
peso	-0.07418	0.05459	9.91059	1.85	0.1869
tiempo	-2.62865	0.38456	250.82210	46.72	<.0001
pe	-0.36963	0.11985	51.05806	9.51	0.0051
pd	-0.02153	0.06605	0.57051	0.11	0.7473
pm	0.30322	0.13650	26.49142	4.93	0.0360

Bounds on condition number: 8.7438, 137.13

Backward Elimination: Step 1

Variable pd Removed: R-Square = 0.8480 and C(p) = 5.1063

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	721.97309	144.39462	27.90	<.0001
Error	25	129.40845	5.17634		
Corrected Total	30	851.38154			

The REG Procedure
Model: MODEL2
Dependent Variable: oxigen
Backward Elimination: Step 1

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	102.20428	11.97929	376.78935	72.79	<.0001
edad	-0.21962	0.09550	27.37429	5.29	0.0301
peso	-0.07230	0.05331	9.52157	1.84	0.1871
tiempo	-2.68252	0.34099	320.35968	61.89	<.0001
pe	-0.37340	0.11714	52.59624	10.16	0.0038
pm	0.30491	0.13394	26.82640	5.18	0.0316

Bounds on condition number: 8.7312, 104.83

Backward Elimination: Step 2
Variable peso Removed: R-Square = 0.8368 and C(p) = 4.8800

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	712.45153	178.11288	33.33	<.0001
Error	26	138.93002	5.34346		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	98.14789	11.78569	370.57373	69.35	<.0001
edad	-0.19773	0.09564	22.84231	4.27	0.0488
tiempo	-2.76758	0.34054	352.93570	66.05	<.0001
pe	-0.34811	0.11750	46.90089	8.78	0.0064
pm	0.27051	0.13362	21.90067	4.10	0.0533

Bounds on condition number: 8.4182, 76.851

All variables left in the model are significant at the 0.1000 level.

The REG Procedure
Model: MODEL2
Dependent Variable: oxigen
Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	pd	5	0.0007	0.8480	5.1063	0.11	0.7473
2	peso	4	0.0112	0.8368	4.8800	1.84	0.1871

En el método *backware* se comenzó introduciendo todas las variables en el modelo, sin y con constante, resultando no significativos algunos parámetros. En los pasos siguientes se van excluyendo variables del modelo hasta obtener como significativo al 90% el mismo modelo que en el método *forward*.

Se concluye que el mejor modelo de ajuste para estos datos es el siguiente:

$$oxigen=98,14-0,19*edad-2,76*tiempo-0,34*pe+0,27*pm$$

Como segundo ejemplo consideramos el modelo *oxigen = a+b*tiempo* para los datos anteriores, que era uno de los modelos que había resultado factible al analizar el método *forward*. Realizamos el ajuste y lo representamos etiquetando las variables de forma adecuada.

```
data salud1;
     set salud;
     label edad      ='edad(anos)'
           peso      ='peso(kg)'
           oxigen    ='oxigeno consumido(ml/kg/min)'
           tiempo    ='1.5 mil tiempo(min)'
           pd              ='pulsaciones en descanso'
           pe        ='pulsaciones en ejercicio'
           pm ='pulsaciones maximas';
   proc reg data=salud1;
     model Oxigen=Tiempo;
     plot Oxigen*Tiempo / cframe=ligr;
   run;
```

La salida es la siguiente:

The REG Procedure
Model: MODEL1
Dependent Variable: oxigen oxigeno consumido(ml/kg/min)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	632.90010	632.90010	84.01	<.0001
Error	29	218.48144	7.53384		
Corrected Total	30	851.38154			

Root MSE	2.74478	R-Square	0.7434
Dependent Mean	47.37581	Adj R-Sq	0.7345
Coeff Var	5.79364		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	82.42177	3.85530	21.38	<.0001
tiempo	1.5 mil tiempo(min)	1	-3.31056	0.36119	-9.17	<.0001

La F para el ajuste global resulta altamente significativa, lo mismo que la T para cada parámetro estimado. Un buen ajuste al 99% será *oxigen = 82,4 – 3,3*tiempo* tal y como se muestra en la salida gráfica de la Figura 2-34.

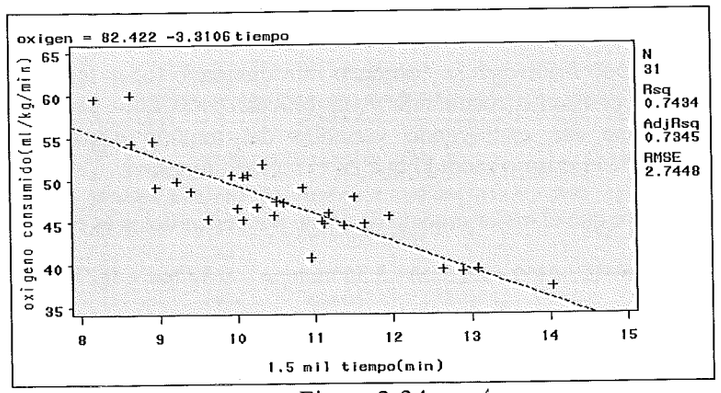


Figura 2-34

Como tercer ejemplo se presenta el ajuste anterior pero sin salida numérica y con salida gráfica representando las franjas de los intervalos de confianza para parámetros estimados y predicciones.

```
data salud1;
     set salud;
     label edad      ='edad(anos)'
           peso      ='peso(kg)'
           oxigen    ='oxigeno consumido(ml/kg/min)'
           tiempo    ='1.5 mil tiempo(min)'
           pd              ='pulsaciones en descanso'
           pe        ='pulsaciones en ejercicio'
           pm ='pulsaciones maximas';

  title 'Intervalos de confianza para parametros y predicciones';
  proc reg data=salud1;
     model Oxigen=Tiempo / noprint;
     plot Oxigen*Tiempo / conf pred cframe=ligr;
  run;
```

La salida se presenta en la Figura 2-35.

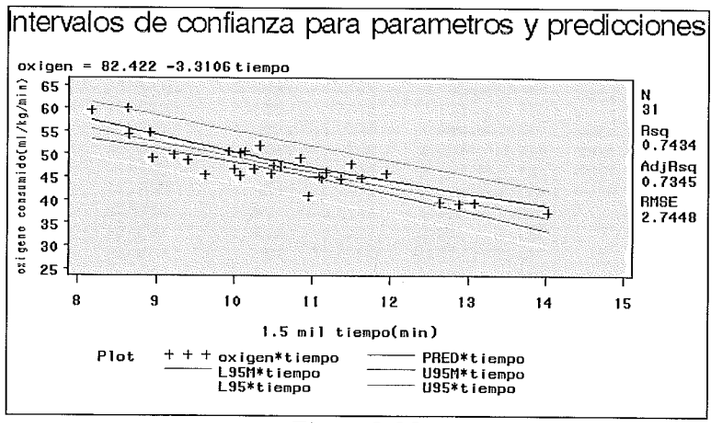


Figura 2-35

Como cuarto ejemplo se considera el modelo *oxigen=a+b*edad+c*peso+ d*tiempo + e*pd + f*pe + g*pm+ h*dif* siendo *dif* la diferencia entre las pulsaciones en ejercicio y en descanso. Ajustar el modelo y realizar un diagnóstico de colinealidad, de autocorrelación y de residuos.

```
title 'Analisis de colinealidad, autocorrelacion y residuos';
data salud2;
   set salud;
   Dif=pe-pd;
proc reg data=salud2;
   model Oxigen=edad peso tiempo pe pm pd dif
   / tol vif collin DW R;
run;
```

La salida es la siguiente:

Analisis de colinealidad, autocorrelacion y residuos

The REG Procedure
Model: MODEL1
Dependent Variable: oxigen

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	722.54361	120.42393	22.43	<.0001
Error	24	128.83794	5.36825		
Corrected Total	30	851.38154			

Root MSE	2.31695	R-Square	0.8487
Dependent Mean	47.37581	Adj R-Sq	0.8108
Coeff Var	4.89057		

NOTE: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.
NOTE: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

Dif = pe - pd

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	102.93448	12.40326	8.30	<.0001	.	0
edad	1	-0.22697	0.09984	-2.27	0.0322	0.66101	1.51284
peso	1	-0.07418	0.05459	-1.36	0.1869	0.86555	1.15533
tiempo	1	-2.62865	0.38456	-6.84	<.0001	0.62859	1.59087
pe	B	-0.36963	0.11985	-3.08	0.0051	0.11852	8.43727
pm	1	0.30322	0.13650	2.22	0.0360	0.11437	8.74385
pd	B	-0.02153	0.06605	-0.33	0.7473	0.70642	1.41559
Dif	0	0	.	.	.	.	.

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation: Intercept	edad	peso	tiempo	pe
1	7.94175	1.00000	0.00001782	0.00011829	0.00015052	0.00016133	2.48828E-14
2	0.02236	18.84725	0.00074589	0.00837	0.01303	0.06437	5.11141E-13
3	0.01647	21.96172	0.00018210	0.30837	0.05936	0.03935	3.80839E-12
4	0.00939	29.08042	0.00150	0.00173	0.61267	0.15240	1.1197E-11
5	0.00862	30.34547	0.00540	0.08366	0.13590	0.61923	2.3832E-12
6	0.00117	82.26016	0.71400	0.50403	0.15810	0.11573	6.59774E-11
7	0.00023674	183.15559	0.27816	0.09372	0.02079	0.00875	4.10021E-10
8	1E-12	2818111	0	0	0	0	1.00000

Collinearity Diagnostics

Number	Proportion of Variation: pm	pd	Dif
1	0.00000486	2.43485E-13	5.25286E-14
2	0.00015694	3.84906E-10	3.27015E-11
3	0.00077401	3.6871E-11	8.07914E-12
4	0.00151	1.09709E-10	2.36231E-11
5	0.00004190	1.91849E-10	3.89728E-11
6	0.00157	4.08367E-10	1.66806E-10
7	0.99595	1.155269E-9	1.274894E-9
8	1.04209E-17	1.00000	1.00000

NOTE: Singularities or near singularities caused grossly large variance calculations. To provide diagnostics, eigenvalues are inflated to a minimum of 1e-12.

The REG Procedure

Model: MODEL1

Dependent Variable: oxigen

Durbin-Watson D	1.711
Number of Observations	31
1st Order Autocorrelation	0.130

The REG Procedure
Model: MODEL1
Dependent Variable: oxigen

Output Statistics

Obs	Dep Var oxigen	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	44.6090	44.4799	0.8734	0.1291	2.146	0.0602	| |	0.000
2	45.3130	48.1954	1.0672	-2.8824	2.057	-1.402	|**| |	0.076
3	54.2970	56.1519	1.1443	-1.8549	2.015	-0.921	|*| |	0.039
4	59.5710	56.8041	1.2075	2.7669	1.977	1.399	|** |	0.104
5	49.8740	51.0710	1.1743	-1.1970	1.997	-0.599	|*| |	0.018
6	44.8110	43.0132	0.6705	1.7978	2.218	0.811	|* |	0.009
7	45.6810	44.8244	1.1907	0.8566	1.988	0.431	| |	0.010
8	49.0910	48.9203	1.0635	0.1707	2.058	0.0829	| |	0.000
9	39.4420	40.2197	1.0438	-0.7777	2.069	-0.376	| |	0.005
10	60.0550	58.0793	1.6245	1.9757	1.652	1.196	|** |	0.198
11	50.5410	48.7762	1.0925	1.7648	2.043	0.864	|* |	0.030
12	37.3880	37.5993	1.5015	-0.2113	1.765	-0.120	| |	0.001
13	44.7540	45.7745	1.0219	-1.0205	2.079	-0.491	| |	0.008
14	47.2730	47.3677	0.8839	-0.0947	2.142	-0.0442	| |	0.000
15	51.8550	46.4703	0.8320	5.3847	2.162	2.490	|**** |	0.131
16	49.1560	50.8615	1.1855	-1.7055	1.991	-0.857	|*| |	0.037
17	40.8360	46.2386	0.6639	-5.4026	2.220	-2.434	|****| |	0.076
18	46.6720	49.3203	0.6103	-2.6483	2.235	-1.185	|**| |	0.015
19	46.7740	47.1135	1.1692	-0.3395	2.000	-0.170	| |	0.001
20	50.3880	47.2738	1.2236	3.1142	1.968	1.583	|*** |	0.138
21	39.4070	39.1567	1.0656	0.2503	2.057	0.122	| |	0.001
22	46.0800	46.4614	1.2187	-0.3814	1.971	-0.194	| |	0.002
23	45.4410	48.8382	0.7598	-3.3972	2.189	-1.552	|***| |	0.041
24	54.6250	54.8806	1.1055	-0.2556	2.036	-0.126	| |	0.001
25	45.1180	44.7887	0.9153	0.3293	2.128	0.155	| |	0.001
26	39.2030	39.1324	1.4192	0.0706	1.831	0.0386	| |	0.000
27	45.7900	45.3528	1.1495	0.4372	2.012	0.217	| |	0.002
28	50.5450	50.7506	1.3230	-0.2056	1.902	-0.108	| |	0.001
29	48.6730	48.4904	1.1986	0.1826	1.983	0.0921	| |	0.000
30	47.9200	46.6774	1.1923	1.2426	1.987	0.626	|* |	0.020
31	47.4670	45.5659	0.7433	1.9011	2.194	0.866	|* |	0.012

Sum of Residuals	0
Sum of Squared Residuals	128.83794
Predicted Residual SS (PRESS)	192.78822

El diagnóstico de colinealidad ha detectado multicolinealidad en el modelo provocada para la variable *dif*. Se observa la presencia de índices de condición superiores a 30. El estadístico de DurbinWatson está cercano a 2 presentándose una leve autocorrelación de orden 1 (ρ=0,13). A primera vista, la tabla de residuos no indica problemas.

Si se utiliza la versión 9.2 de SAS con la misma sintaxis de este ejemplo precedida de la activación del módulo gráfico avanzado mediante la línea de código:

```
ods graphics on;
```

tendremos la salida gráfica adicional de las figuras 2-36 (heteroscedasticidad, linealidad, normalidad residual y puntos influyentes sin problemas) y 2-37 (heteroscedastcidad):

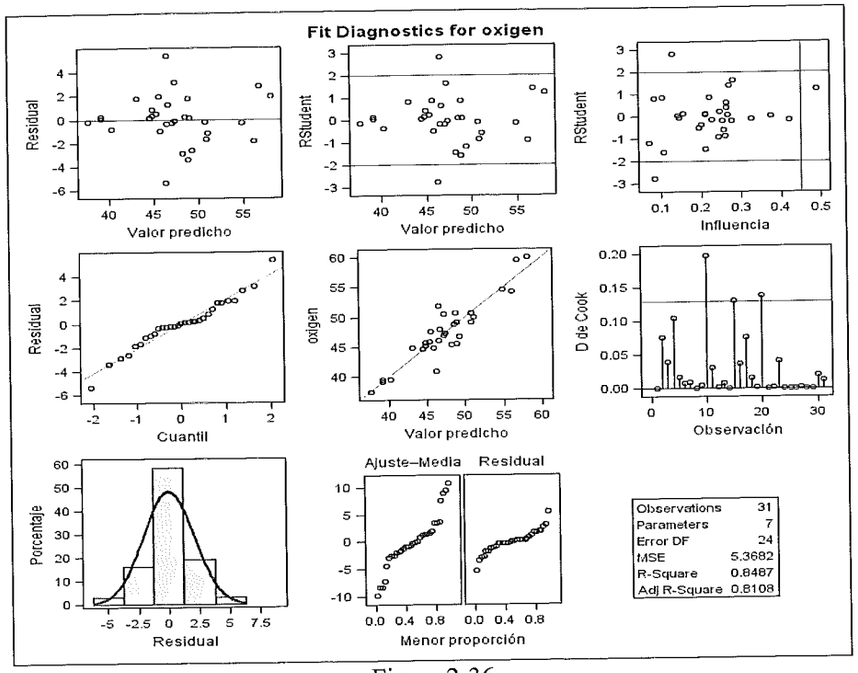


Figura 2-36

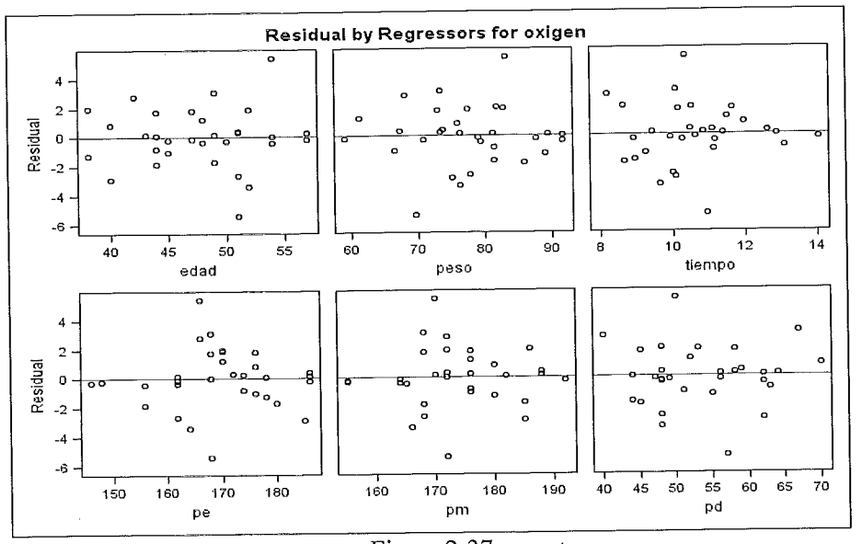


Figura 2-37

Como quinto ejemplo se considera el modelo *oxigen=a+b*tiempo* y se realiza un diagnóstico gráfico de normalidad de los residuos.

```
proc reg data=salud;
title 'PP Plot';
   model Oxigen=tiempo / noprint;
   plot npp.*r.
;
run;
```

La salida se presenta en la figura 2-38 que indica que los residuos tienen una distribución normal, ya que los puntos del gráfico de probabilidades P-P se ajustan muy bien a la diagonal del primer cuadrante.

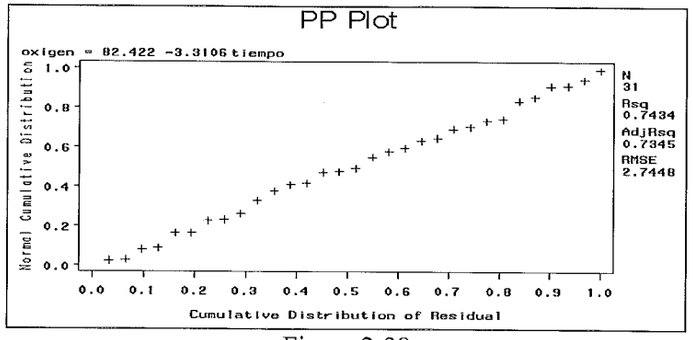


Figura 2-38

Como quinto ejemplo realizamos un análisis de la autocorrelación en el modelo *oxigen=a+b*edad + b*tiempo + c*pe + d*pm* a través del gráfico de residuos contra número de observación. Asimismo realizaremos un análisis gráfico de la heteroscedasticidad representando residuos contra valores predichos y valores observados contra valores predichos. Por último, analizaremos la linealidad y la heteroscedasticidad graficando residuos estudentizados contra variables independientes.

```
proc reg data=salud;
   model Oxigen=edad tiempo pe pm;

   /*Analisis grafico de la autocorrelacion*/

plot residual.*obs.;

   /*Analisis grafico de la heteroscedasticidad*/

plot (residual. obs.)*predicted.;

   /*Analisis grafico de la linealidad*/

plot (student.)*(edad tiempo pe pm);
run;
```

La salida gráfica, que se presenta en las Figuras 2-39 a 2-45, no detecta ningún problema de autocorrelación, ni de heteroscedasticidad, ni de falta de linealidad, ya que todos los gráficos presentan una distribución aleatoria de sus puntos. La mayoría de estos gráficos se obtenían directamente con la versión 9.2 de SAS (Figuras 2-36 y 2-37).

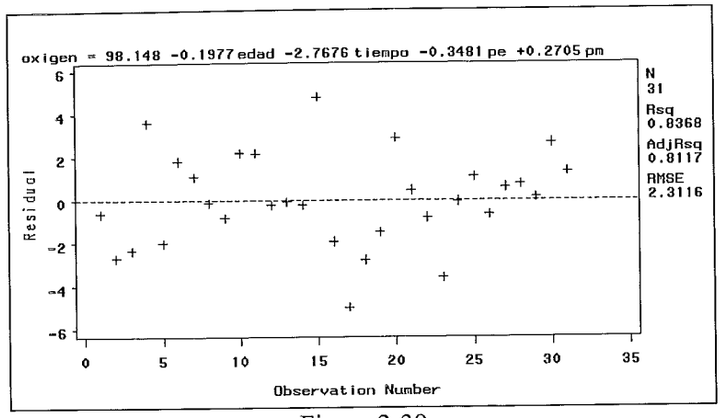


Figura 2-39

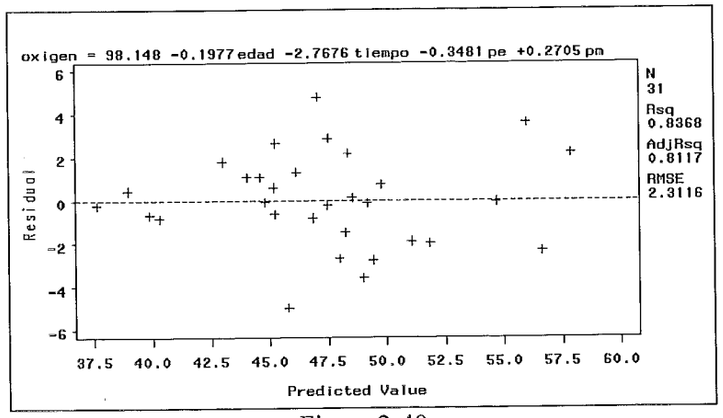


Figura 2-40

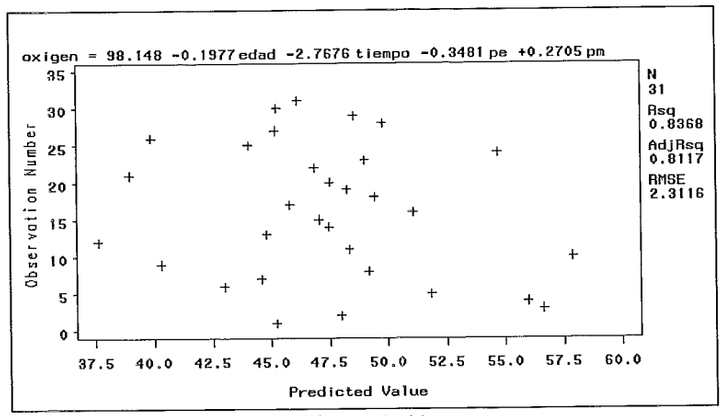


Figura 2-41

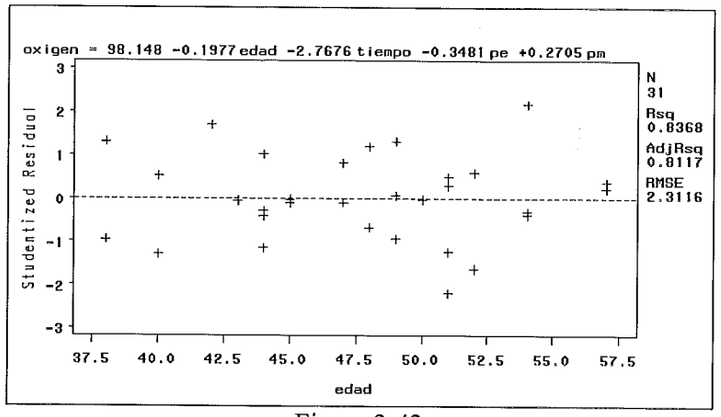


Figura 2-42

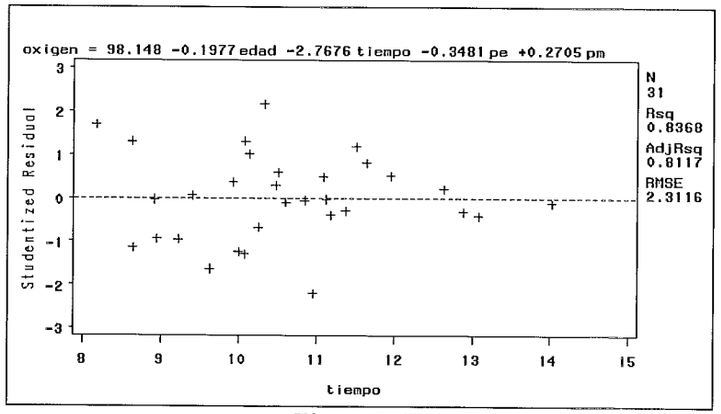


Figura 2-43

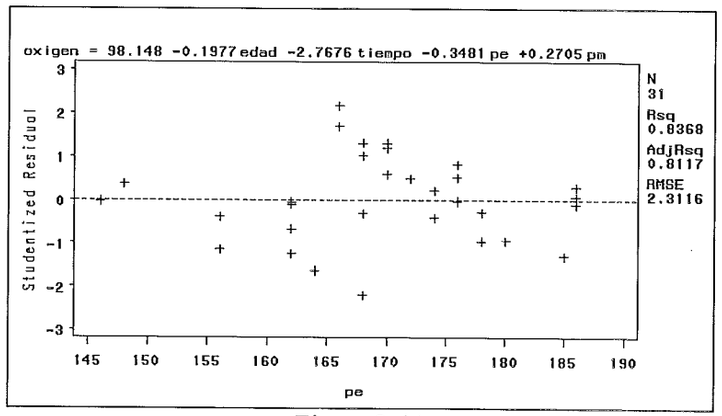


Figura 2-44

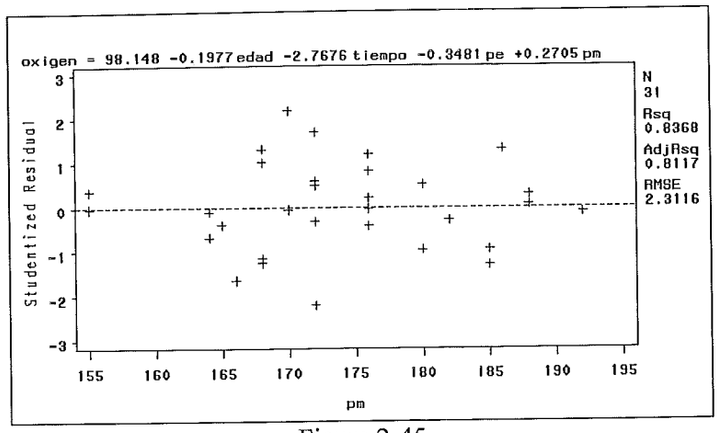


Figura 2-45

La salida numérica del ajuste del modelo es la siguiente:

The REG Procedure
Model: MODEL1
Dependent Variable: oxigen

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	712.45153	178.11288	33.33	<.0001
Error	26	138.93002	5.34346		
Corrected Total	30	851.38154			

Root MSE	2.31159	R-Square	0.8368
Dependent Mean	47.37581	Adj R-Sq	0.8117
Coeff Var	4.87927		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	98.14789	11.78569	8.33	<.0001
edad	1	-0.19773	0.09564	-2.07	0.0488
tiempo	1	-2.76758	0.34054	-8.13	<.0001
pe	1	-0.34811	0.11750	-2.96	0.0064
pm	1	0.27051	0.13362	2.02	0.0533

Se observa una buena significatividad global del modelo, ya que el contraste de la F de Fisher Snedokor presenta un p-valor menor que 0,05. A su vez el coeficiente de determinación ajustado tiene un valor muy alto (0,83). Por su parte, los coeficientes individuales también son significativos al 94% ya que el p-valor del contraste de a T de Student de todos ellos es menor que 0,06.

Como sexto ejemplo, vamos a contrastar en el modelo anterior las hipótesis de nulidad conjunta de los parámetros *edad* y *pm*, y también contrastaremos la hipótesis *edad+pe-2*pm=0*.

```
proc reg data=salud;
   model Oxigen=edad tiempo pe pm;

   /*Contraste de nulidad*/

nulidad: test edad=0, pm=0;

   /*Contraste para la combinacion lineal*/

clineal: test edad+pe-2*pm=0;
run;
```

La salida es la siguiente:

The REG Procedure
Model: MODEL1

Test nulidad Results for Dependent Variable oxigen

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	32.09467	6.01	0.0072
Denominator	26	5.34346		

The REG Procedure
Model: MODEL1

Test clineal Results for Dependent Variable oxigen

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	47.20229	8.83	0.0063
Denominator	26	5.34346		

Tanto el test de nulidad conjunta de los dos parámetros, como el test de la combinación lineal indica que se pueden aceptar ambas hipótesis nulas al 99% de confianza, ya que los p-valores de los dos contrastes de la F son menores que 0,01.

Como séptimo ejemplo, vamos a estimar el modelo anterior sujeto a las restricciones *edad + pm = 0* y *edad + pe = 2*pm=0*. Estas restricciones son lógicas en el modelo debido al resultado del contrastre anterior.

```
proc reg data=salud;
   model Oxigen=edad tiempo pe pm;

   /*Ajuste con restricciones*/

restrict pe+pm=0, edad+pe=2*pm;
run;
```

La salida es la siguiente:

```
                    The REG Procedure

                    Model: MODEL1

              Dependent Variable: oxygen

NOTE: Restrictions have been applied to parameter estimates.

                    Analysis of Variance

                              Sum of          Mean
Source                 DF    Squares        Square    F Value    Pr > F

Model                   2  638.83776     319.41888      42.08    <.0001
Error                  28  212.54378       7.59085
Corrected Total        30  851.38154


          Root MSE              2.75515    R-Square     0.7504
          Dependent Mean       47.37581    Adj R-Sq     0.7325
          Coeff Var             5.81552

                    Parameter Estimates

                         Parameter      Standard
Variable      DF          Estimate         Error    t Value    Pr > |t|

Intercept      1          86.25334       5.80898      14.85      <.0001
edad           1          -0.08619       0.09745      -0.88      0.3840
tiempo         1          -3.27313       0.36502      -8.97      <.0001
pe             1           0.02873       0.03248       0.88      0.3840
pm             1          -0.02873       0.03248      -0.88      0.3840
RESTRICT      -1        -224.24363     120.21785      -1.87      0.0607*
RESTRICT      -1         -51.72068      17.97431      -2.88      0.0023*

          * Probability computed using beta distribution.
```

El ajuste global del modelo es correcto al 95% (p-valor de la F menor que 0,05 y coeficiente de determinación ajustado alto = 0,75), pero los parámetros *edad, pe* y *pm* no resultan significativos según el contraste de la T, ya que sus p-valores son muy superiores a 0,05.

Habría que redefinir las restricciones del modelo o tomar cualquier otra solución alternativa de inclusión o eliminación de variables.

STATA Y EL TRABAJO BÁSICO CON EL MODELO DE REGRESIÓN MÚLTIPLE

El software STATA utiliza el comando *regress* para ajustar el modelo lineal de regresión simple. La sintaxis básica de este comando es la siguiente:

Regress $Y\ X_1\ X_2 \ldots X_n$

donde Y es la variable dependiente y X_i son las variables independientes.

Como ejemplo consideramos el fichero *auto.dta* con datos sobre automóviles que, entre otras, contiene las variables *mpg* (consumo), *weight* (peso) y *foreign* (vale uno para coches extranjeros y vale cero para los coches nacionales). Se trata de estimar el modelo:

$$mpg = \beta_0 + \beta_1 weight + \beta_2 weight^2 + \beta_3 foreign + u$$

La sintaxis a escribir en STATA sería la siguiente:

```
. use auto

. generate weightsq=weight^2

. regress mpg weight weightsq foreign
```

La salida se presenta a continuación:

```
      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  3,    70) =   52.25
       Model |  1689.15372     3  563.05124            Prob > F      =  0.0000
    Residual |   754.30574    70  10.7757963           R-squared     =  0.6913
-------------+------------------------------           Adj R-squared =  0.6781
       Total |  2443.45946    73  33.4720474           Root MSE      =  3.2827

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0165729   .0039692    -4.18   0.000    -.0244892   -.0086567
    weightsq |   1.59e-06   6.25e-07     2.55   0.013     3.45e-07    2.84e-06
     foreign |    -2.2035   1.059246    -2.08   0.041      -4.3161   -.0909002
       _cons |   56.53884   6.197383     9.12   0.000     44.17855    68.89913
------------------------------------------------------------------------------
```

Se observa que los parámetros estimados son significativos individualmente al 95% ya que los p-valores son menores que 0,05. También existe fuerte significatividad conjunta de los parámetros porque el p-valor de la F es prácticamente nulo. Adicionalmente se muestran intervalos de confianza al 95% para los parámetros. El R^2 ajustado nos indica que el modelo explica el 67,81% de la variabilidad.

El modelo estimado es el siguiente:

$$mpg = 56.53884 - 0.0165729 weight + 0{,}00000159 weight^2 - 2{,}2035\, foreign + u$$

La regresión anterior sin constante se realizaría como sigue:

```
. regress mpg weight weightsq foreign, noconstant

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  3,    71) =  492.45
       Model |  34356.8305     3  11452.2768           Prob > F      =  0.0000
    Residual |  1651.16952    71  23.2559087           R-squared     =  0.9541
-------------+------------------------------           Adj R-squared =  0.9522
       Total |       36008    74  486.594595           Root MSE      =  4.8224
```

```
------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .0189537   .0011281    16.80   0.000     .0167043     .021203
    weightsq |  -3.78e-06   3.09e-07   -12.22   0.000    -4.39e-06   -3.16e-06
     foreign |   1.811004   1.415473     1.28   0.205    -1.01137     4.633377
------------------------------------------------------------------------------
```

A continuación se utiliza el comando *test* para presentar ejemplos sobre contrastes de restricciones en los parámetros del modelo. Primero contrastaremos $\beta_1=0$ y después $\beta_3 = 2\beta_1$. En ambos casos el p-valor menor que 0,05 indica la aceptación de las hipótesis al menos al 95% de confianza.

```
test weightsq = 0

 ( 1)  weightsq = 0

       F(  1,    70) =    9.12
            Prob > F =    0.0035

test foreign=2*weight

 ( 1) - 2 weight + foreign = 0

       F(  1,    70) =    4.44
            Prob > F =    0.0386
```

Podemos graficar residuos contra valores predichos con la línea horizontal en el origen (Figura 2-46) mediante la sintaxis siguiente:

```
rvfplot, yline(0)
```

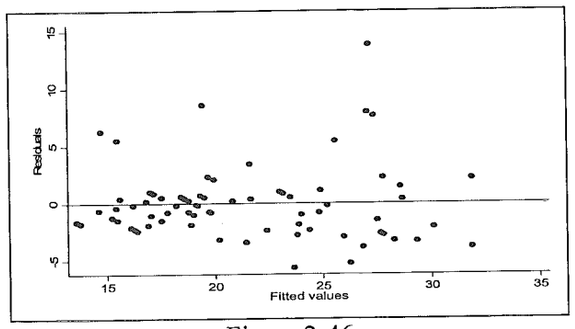


Figura 2-46

STATA Y EL TRABAJO BÁSICO CON EL MODELO DE REGRESIÓN MÚLTIPLE A TRAVÉS DE MENÚS

Como ejemplo ilustrativo supongamos que un analista comercial está pensando en adquirir un grupo de pequeños edificios de oficinas en un distrito comercial conocido. El analista puede utilizar el análisis de regresión lineal múltiple para estimar el valor de un edificio de oficinas Y en un área determinada basándose en las variables X_1, X_2, X_3 y X_4 siguientes.

Variable	Explicación
Y	Valor tasado del edificio de oficinas
X_1	Superficie en metros cuadrados
X_2	Número de oficinas
X_3	Número de entradas
X_4	Antigüedad del edificio en años

El analista elige al azar una muestra de 11 edificios de oficinas de 1500 edificios posibles, y obtiene los datos de la Figura 2-47 recogidos en el fichero en formato Stata de nombre *regres1.dta*.

X1	X2	X3	X4	Y
2310	2	2	20	142.000
2333	2	2	12	144.000
2356	3	1,5	33	151.000
2379	3	2	43	150.000
2402	2	3	53	139.000
2425	4	2	23	169.000
2448	2	1,5	99	126.000
2471	2	2	34	142.900
2494	3	3	23	163.000
2517	4	4	55	169.000
2540	2	3	22	149.000

Figura 2-47

Para ajustar un modelo de regresión lineal con Stata a través de menús se utiliza la ruta *Statistics* → *Linear Models and Related* → *Linear Regressión* (Figura 2-48) y se rellena la pantalla de entrada como se indica en la Figura 2-49.

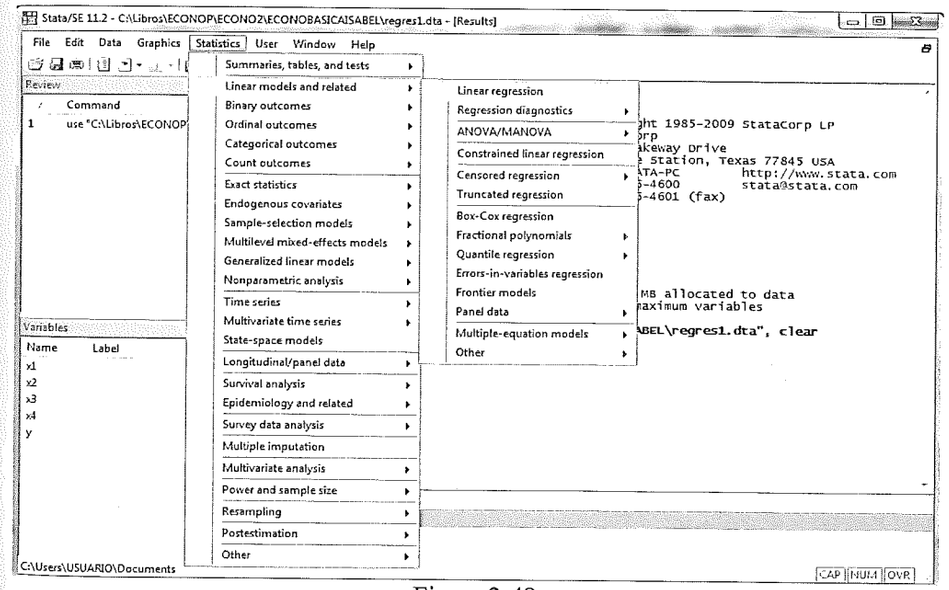


Figura 2-48

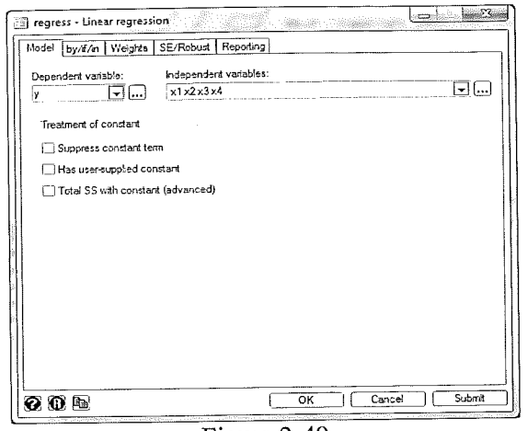


Figura 2-49

Al pulsar *Ok*, se obtiene la salida numérica que se muestra a continuación, que incluye estadísticos de regresión, cuadro del análisis de la varianza del modelo, estimadores, contrastes de significatividad de F y de T con sus *p*-valores asociados (todos ellos muy pequeños indicando la fuerte significatividad individual y conjunta de los parámetros estimados), intervalos de confianza para los parámetros estimados al 95% de confianza.

```
. use "C:\Libros\regres1.dta", clear

. regress y x1 x2 x3 x4

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  4,     6) =  459.75
       Model |  1.7324e+09     4  433098330           Prob > F      =  0.0000
    Residual |  5652135.32     6  942022.553           R-squared     =  0.9967
-------------+------------------------------           Adj R-squared =  0.9946
       Total |  1.7380e+09    10  173804545           Root MSE      =  970.58

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   27.64139   5.429374     5.09   0.002     14.35619    40.92659
          x2 |   12529.77   400.0668    31.32   0.000     11550.84     13508.7
          x3 |   2553.211   530.6692     4.81   0.003      1254.71    3851.711
          x4 |  -234.2372   13.26801   -17.65   0.000    -266.7028   -201.7715
       _cons |   52317.83   12237.36     4.28   0.005     22374.09    82261.58
------------------------------------------------------------------------------
```

Para realizar una diagnosis de la regresión más avanzada, se utiliza la ruta la ruta *Statistics* → *Linear Models and Related* → *Regression diagnostics* (Figura 2-50). La opción *Specification test, etc.* Realiza contrastes avanzados de autocorrelación, heteroscedasticidad, etc. que serán tratados en profundidad en próximos capítulos. La opción *Added variable plot* realiza un gráfico de cada regresor (todos o los especificados en el campo *Variables* de la Figura 2-51) contra la variable dependiente (Figura 2-52) con el objeto de ver cómo influye cada regresor sobre la citada variable dependiente. La opción *Component-plus-residual-plot* grafica la variable independiente especificada en el campo variable de la Figura 2-53 contra los valores predichos (Figura 2-54).

La opción *Residual versus fitted plot* grafica los residuos contra los valores predichos (Figura 2-55). El gráfico debe ser aleatorio para que le modelo no tenga problemas de heteroscedasticidad o falta de linealidad. La opción *Residual versus predictor plot* grafica la variable independiente introducida en el campo *Variable* de la Figura 2-56 contra los residuos (Figura 2-57). Este gráfico debe de ser aleatorio para cada variable dependiente, en cuyo caso no habrá heteroscedaticidad. Si para alguna variable este gráfico no es aleatorio, dicha variable será culpable de heteroscedasticidad en el modelo.

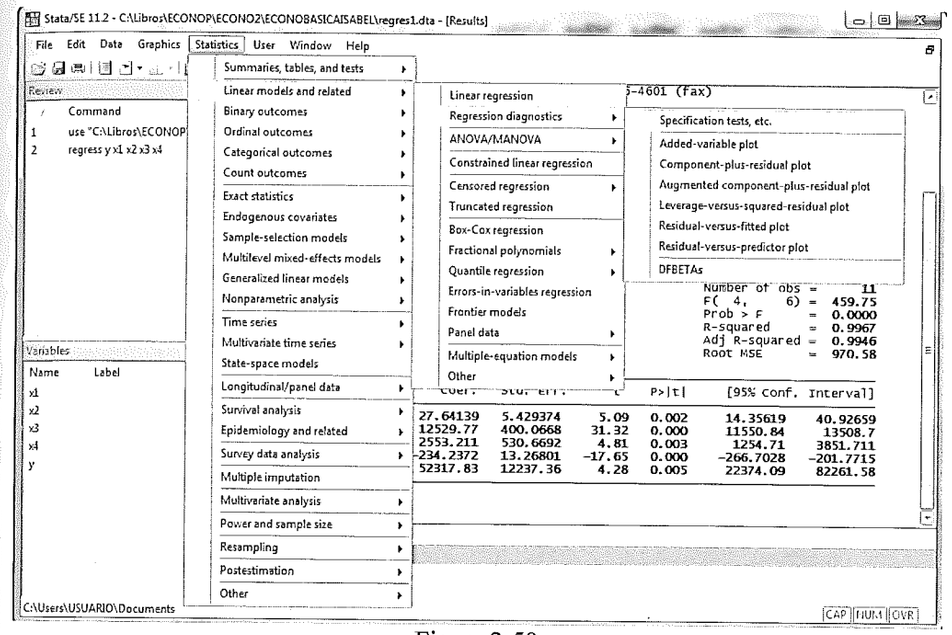


Figura 2-50

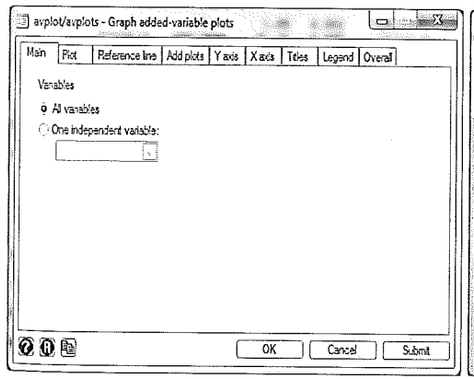


Figura 2-51

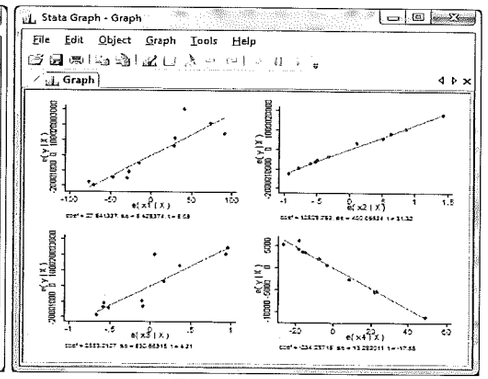


Figura 2-52

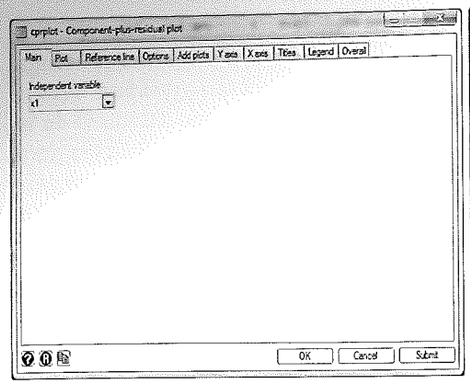


Figura 2-53

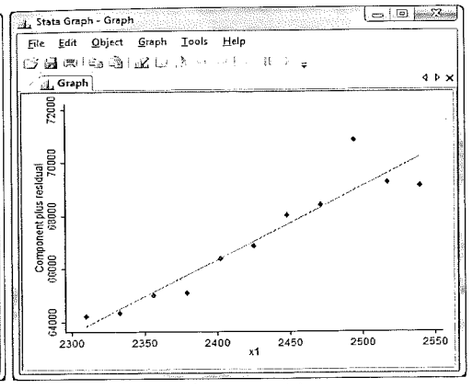


Figura 2-54

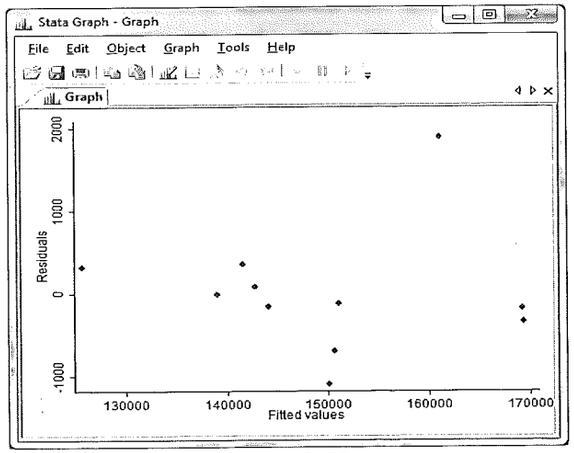


Figura 2-55

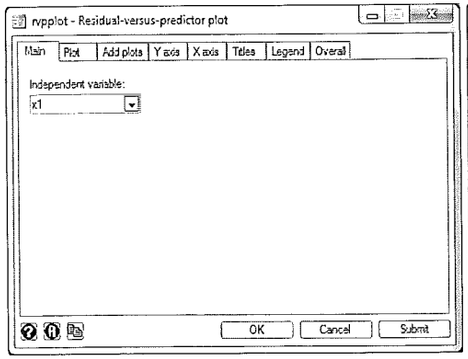


Figura 2-56

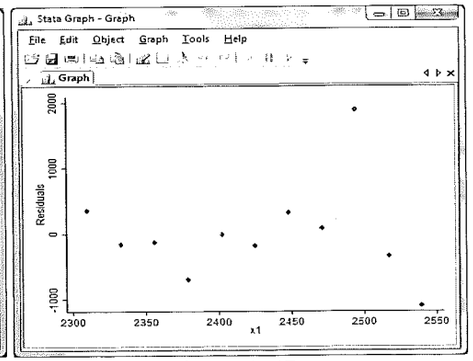


Figura 2-57

Ejercicio 2-1. El archivo en formato Eviews 2-1.wk1 contiene los datos de las variables Y, X1, X2, X3 y X4. Se trata de analizar la relación existente entre Y como variable dependiente y X1, X2, X3 y X4 como variables independientes. Una vez hallada la relación pedida entre las variables, evaluar la capacidad predictiva del modelo y hallar predicciones de Y para los valores siguientes de las variables (X1, X2, X3 y X4): (302, 9, 44, 42), (351, 8, 65, 62) y (381, 9, 52, 53).

Realizaremos el ejercicio mediante la herramienta Eviews. Podemos iniciar el problema realizando un análisis conjunto de las series con la finalidad de ver las relaciones entre ellas. Para realizar esta tarea con Eviews, se carga el fichero de trabajo (2-1.*wk1*) usando *File* → *Open* → *Eviews Workfile*... y se seleccionan todas las variables del modelo (con la tecla *Ctrl.* presionada se eligen todas las variables con el ratón). A continuación se hace clic con el botón derecho sobre la selección y se elige *Open* → *as Group* (Figura 2-58). En la pantalla *Group* se elige *View* → *Multiple Graphs* → *Scatter* → *First series against all*... (Figura 2-59) para obtener la Figura 2-60, que permite intuir una relación lineal positiva entre la variable dependiente y cada una de las variables independientes. Esta relación es más fuerte con X_1 y X_3. Podemos cuantificar las relaciones entre las variables mediante su matriz de correlaciones (en el menú *View* de la pantalla *Group* en la Figura 2-59 se elige *Correlations* → *Pairwise Simples*) de la Figura 2-61. Se observa que X_1 y X_3 son las variables que presentan mayor correlación con Y y que existe una cierta relación entre el X_1 y X_2. También hay relación entre esta última variable y X_3. Es lógico entonces realizar un ajuste lineal de la variable Y respecto de las demás variables.

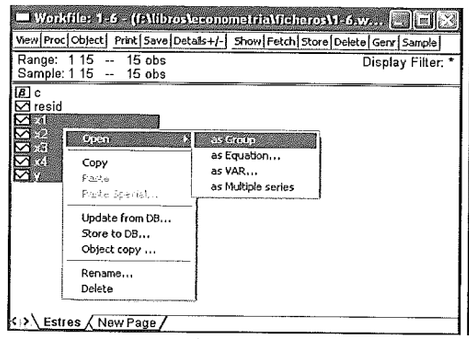


Figura 2-58

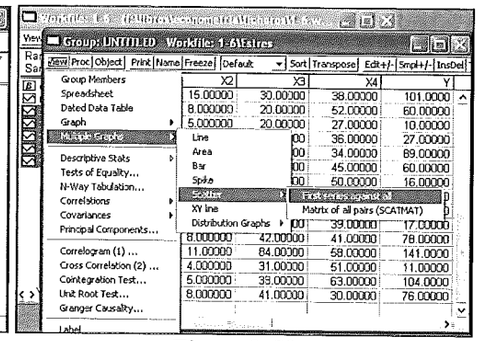


Figura 2-59

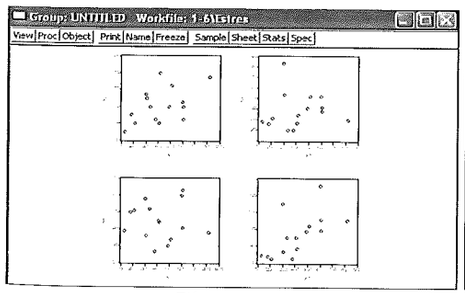
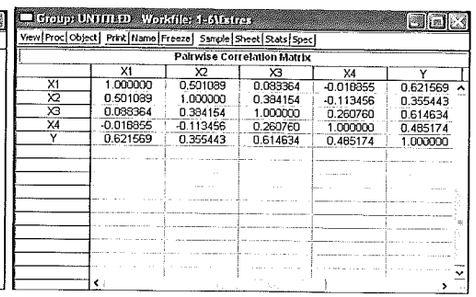


Figura 2-60



	X1	X2	X3	X4	Y
X1	1.000000	0.501089	0.088364	-0.018855	0.621569
X2	0.501089	1.000000	0.384154	-0.113456	0.355443
X3	0.088364	0.384154	1.000000	0.260760	0.614634
X4	-0.018855	-0.113456	0.260760	1.000000	0.485174
Y	0.621569	0.355443	0.614634	0.485174	1.000000

Figura 2-61

Para realizar el ajuste lineal con Eviews, se elige *Quick* → *Estimate Equation*. Se rellena la pantalla *Equation Specification* de la solapa *Specification* como se indica en la Figura 2-62, se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar* (Figura 2-63). Se obtienen los resultados de la Figura 2-64. Se observa que todos los parámetros estimados (salvo el coeficiente de X_2) son significativamente distintos de cero al 95% (p-valores menores que 0,05) y que la significatividad conjunta es muy alta (p-valor de la F muy pequeño). El R^2 y el R^2 ajustado son 0,84 y 0,78 respectivamente (muy altos) con lo que la variabilidad explicada es alta. La desviación típica estimada del error es 24,03 y los criterios de información de Akaike y Schwartz tienen valores pequeños con lo que la capacidad explicativa del modelo es buena. El estadístico de Durban Watson tiene un valor no demasiado alejado de 2, lo que indica ausencia de posibles problemas de autocorrelación. Si sobre la pantalla *Equation* elegimos *View* → *Representation* se obtiene la ecuación de ajuste del modelo (Figura 2-65).

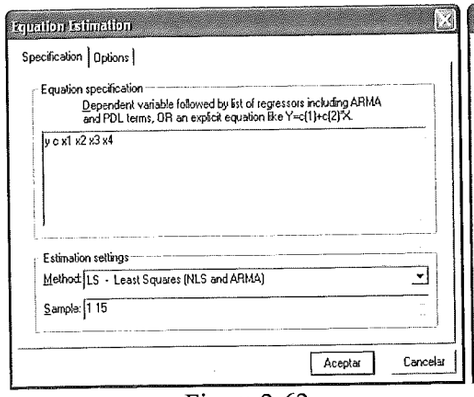


Figura 2-62

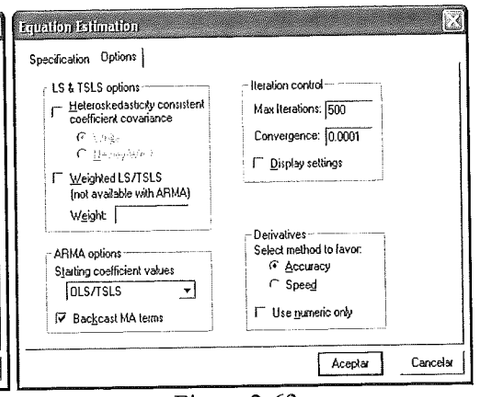


Figura 2-63

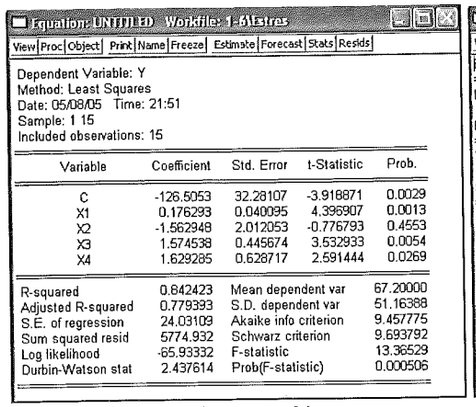


Figura 2-64

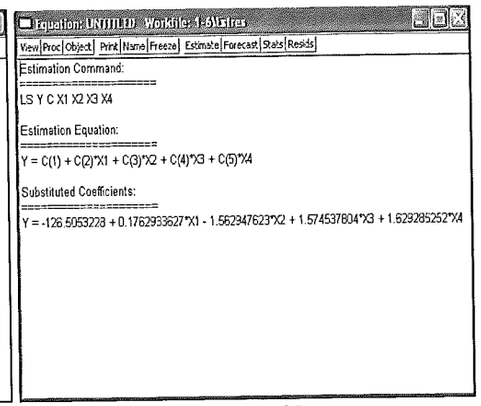


Figura 2-65

Se observa que un aumento en una unidad de la variable X_1 provoca un aumento en Y de 0,176 unidades suponiendo constantes el resto de las variables. Un aumento de una unidad en la variable X_3 con las demás variables constantes provoca un aumento en Y de 1,57 unidades. Para la variable X_4 el aumento es superior (1,63). La variable X_2 debe eliminarse del modelo por no haber resultado significativamente distinta de cero. Para ello, utilizamos la opción *Proc* → *Specify/Estimate* (Figura 2-66) y rellenamos la pantalla *Equation Estimation* sin la variable X_2 (Figura 2-67). Al hacer clic en *Aceptar* se obtiene la estimación del modelo (Figura 2-68). Ahora todas las variables del modelo han aumentado su significatividad (p-valores más pequeños) y los valores de R^2 son mejores (más altos). La significatividad conjunta de los parámetros del modelo también ha mejorado (p-valor de la F más pequeño). La opción *View* → *Representations* muestra el nuevo modelo estimado (Figura 2-69).

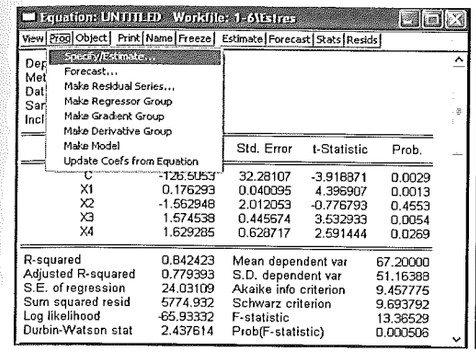


Figura 2-66

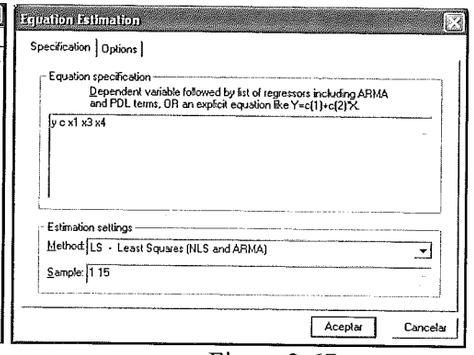


Figura 2-67

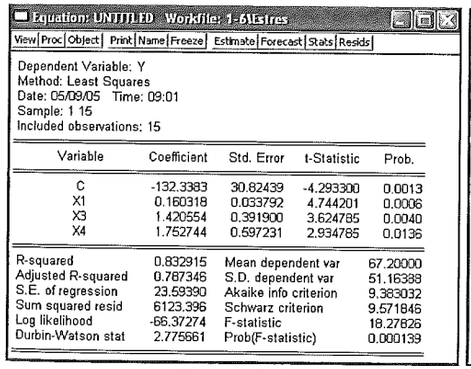


Figura 2-68

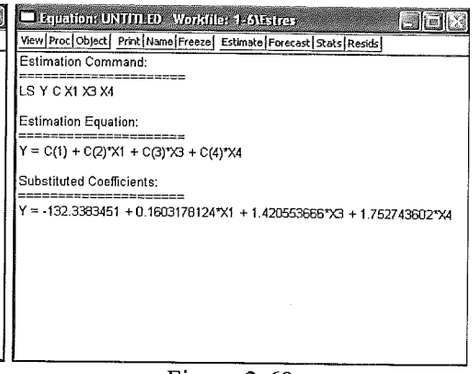


Figura 2-69

Se observa que un aumento en una unidad de la variable X_1 provoca un aumento en Y de 0,16 unidades suponiendo constantes el resto de las variables. Un aumento de una unidad en la variable X_3 con las demás variables constantes provoca un aumento en Y de 1,42 unidades. Para la variable X_4 el aumento es superior (1,75).

Se ve que al eliminar del modelo la variable X_2, los efectos de las variables independientes sobre la variable dependiente han disminuido levemente. Por tanto, la introducción de la variable no significativa en el modelo provocaba una sobreestimación de los efectos de las restantes variables sobre la variable dependiente.

Una vez estimado el modelo es muy conveniente realizar un *análisis de los residuos* para detectar posibles anomalías. Para ello, en la pantalla *Equation*, se utilizan las opciones de *Actual, Fitted, Residual* del menú *View* (Figura 2-70). La subopción *Actual, Fitted, Residual Table* ofrece los resultados de la Figura 2-71. Se obtienen los valores de la variable endógena observados Y_t y estimados $\hat{Y}_t$ y el valor de los residuos. El gráfico de los residuos presenta dos líneas discontinuas situados a una distancia de $\pm 2S$ de la línea central continua, siendo S el error estándar de la regresión (estimador de la desviación típica del error). Estas bandas permiten detectar la presencia de residuos significativamente distintos de cero (atípicos). En nuestro caso, sólo el residuo número 8 se sale de las bandas (aunque no demasiado). La subopción *Actual, Fitted, Residual Graph* muestra el gráfico de la Figura 2-72 que representa los residuos en el eje de ordenadas de la izquierda y los valores de la variable endógena observados y estimados en el eje de ordenadas de la derecha. De esta forma, en la parte superior del gráfico aparecen las gráficas de los valores de la variable endógena observados y estimados, que han de coincidir lo más posible. En la parte inferior aparece el gráfico de los residuos con las bandas $\pm 2S$, que coincide con el gráfico de la Figura 2-71. Este gráfico ha de ser lo más aleatorio posible sin comportamiento sistemático de ningún tipo para evitar la existencia de algún error de especificación en el modelo o la presencia de otros problemas como la autocorrelación.

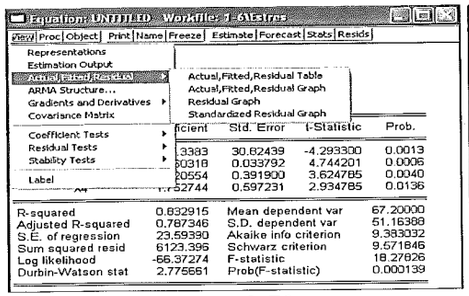


Figura 2-70

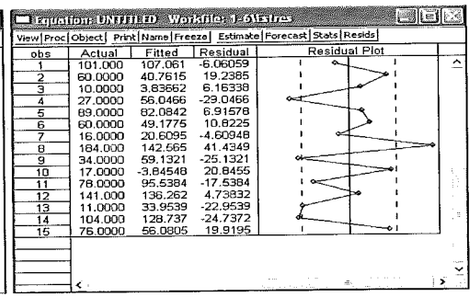


Figura 2-71

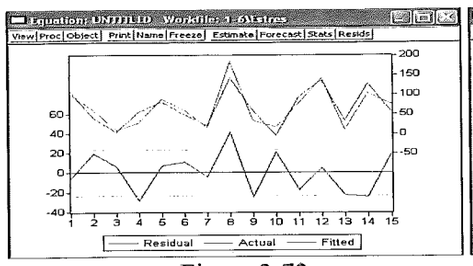


Figura 2-72

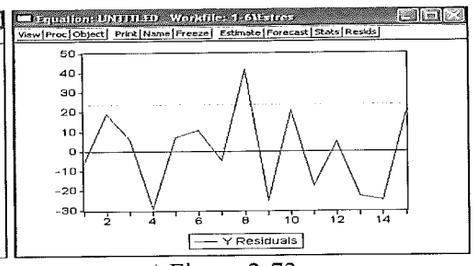


Figura 2-73

La opción *Residual Graph* de la Figura 2-70 presenta solamente el gráfico de los residuos y la opción *Standardized Residual Graph* presenta el gráfico de los residuos estandarizados (Figura 2-73). No se observan problemas de comportamiento sistemático residual por lo que puede suponerse aleatoriedad de los residuos.

Para *calcular predicciones* (tres en nuestro caso), nos situamos en la pantalla *Workfile* y elegimos *Proc → Structure/Resize Current Page* (Figura 2-74) para ampliar el tamaño de la muestra hasta el valor 18 (Figura 2-75). Al pulsar *OK* se obtiene el aviso de la Figura 2-76 indicativo de la inserción de una nueva observación en la muestra (que de momento no tiene valores). Si es necesario, con *Proc → Set Sample...* (Figura 2-84) se amplía el tamaño de la muestra a 18 rellenando la pantalla *Sample* como se indica en la Figura 2-77. La siguiente tarea es introducir los valores de Y, X_1, X_2, X_3 y X_4 para los que se obtendrá la predicción. Para ello, seleccionamos como grupo las cinco variables eligiendo *Open → as Group* en el menú emergente obtenido al hacer clic con el botón derecho del ratón sobre las cuatro variables seleccionadas simultáneamente (Figura 2-78). A continuación, en la pantalla *Group* se pulsa en *Edit* y se introducen los valores de las variables (Figura 2-79). La tarea siguiente es elegir *Proc → Forecast* en la pantalla *Equation* (Figura 2-80) y rellenar la pantalla *Forecast* como se indica en la Figura 2-81. (El vector *Yf* contiene las predicciones y *SEf* sus errores estándar). Al pulsar *OK* se obtiene el gráfico de la Figura 2-82 con los estadísticos de predicción. Además, en la pantalla *Workfile* aparecen los dos vectores *Yf* y *SEf* en la pantalla Workfile (Figura 2-83).

Al hacer clic sobre *yf* en la Figura 2-83 se obtiene la Figura 2-84, cuyos últimos valores son las predicciones de la variable dependiente pedidas.

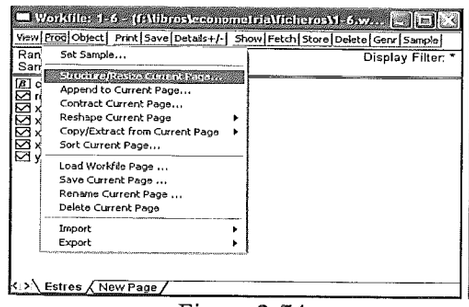


Figura 2-74

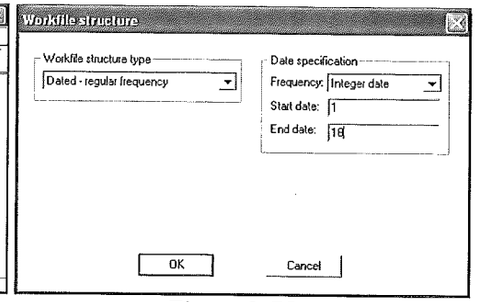


Figura 2-75

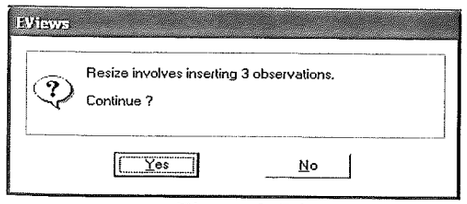


Figura 2-76

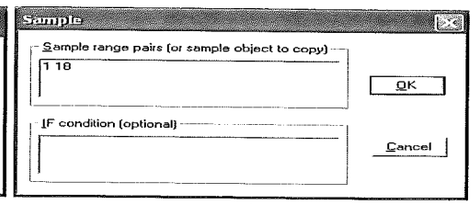


Figura 2-77

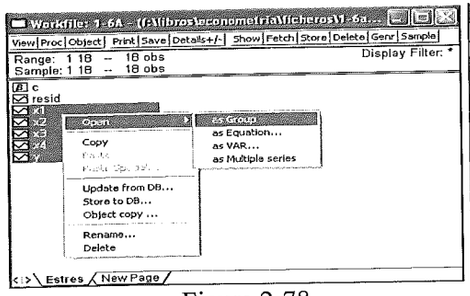


Figura 2-78

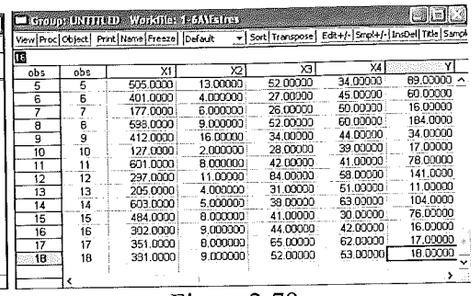


Figura 2-79

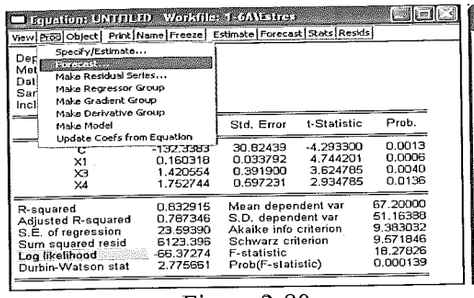


Figura 2-80

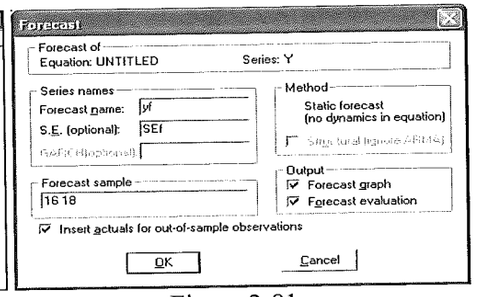


Figura 2-81

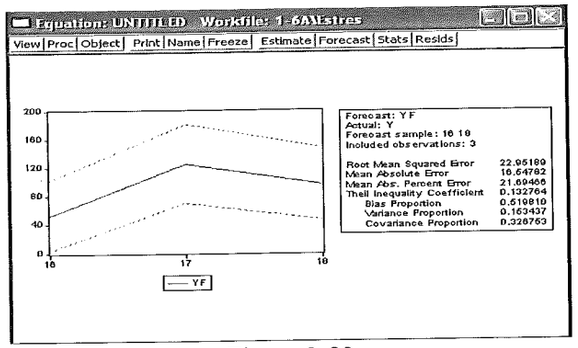


Figura 2-82

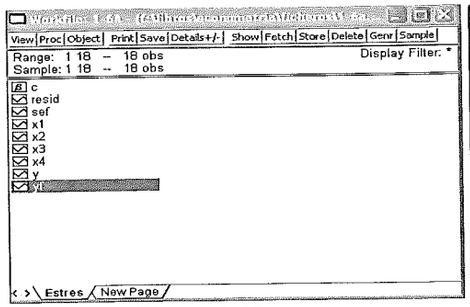


Figura 2-83

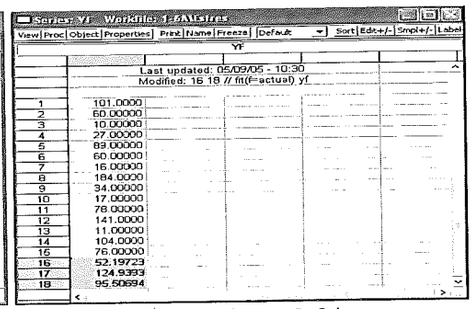


Figura 2-84

Para evaluar la capacidad predictiva del modelo con Eviews, se utilizan los estadísticos de la Figura 2-82. Cuanto más próximos estén a cero los valores de los cuatro primeros estadísticos, mejor será la capacidad predictiva del modelo, lo que permitirá comparar un modelo con otros alternativos. Las tres proporciones varían entre cero y uno y también es conveniente que sean pequeñas. Los resultados obtenidos para nuestro modelo en la Figura 2-82 muestran un coeficiente de Theil muy cercano a cero, aunque los otros estadísticos no sean demasiado pequeños. La proporción de la varianza tiene un valor pequeño y la proporción de la covarianza tiene un valor más o menos aceptable, aunque no ocurre lo mismo con la proporción del sesgo, que tiene un valor algo alto. No olvidemos que sólo tenemos tres observaciones adicionales para la predicción, lo que no permite alcanzar una capacidad de predicción demasiado buena.

Ejercicio 2-2. Consideramos el modelo $y_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u_t$ y los datos del archivo en formato Eviews 2-2.wk1. Estimar el modelo y calcular los coeficientes de determinación y de determinación corregido. Realizar también los contrastes de hipótesis: $\beta_0=\beta_1=\beta_2=0$ (contraste de significatividad conjunta de los parámetros del modelo), $\beta_1=\beta_2=0$, $\beta_1=10\beta_2$, $2\beta_0+2\beta_1+7\beta_2=50$, ($\beta_1=10\beta_2$, $2\beta_0+2\beta_1+7\beta_2=50$). Realizar también el ejercicio a través de los datos del archivo en formato STATA 2-2.dta.

Para resolver este problema con Eviews, se carga el fichero de trabajo (2-2.*wk1*) usando *File* → *Open* → *Eviews Workfile...* y se elige *Quick* → *Estimate Equation* (Figura 2-85). Se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar* (Figura 2-86). Se obtienen los resultados de la Figura 2-87. Se observa que la constante tiene una significatividad inferior al 90% (es distinta de cero al 90% de confianza), el coeficiente de X_1 es muy significativamente distinto de cero (p-valor muy pequeño) y el coeficiente de X_2 no es significativamente distinto de cero (p-valor muy alto). El R^2 y el R^2 ajustado son 0,973 y 0,962 respectivamente (muy altos).

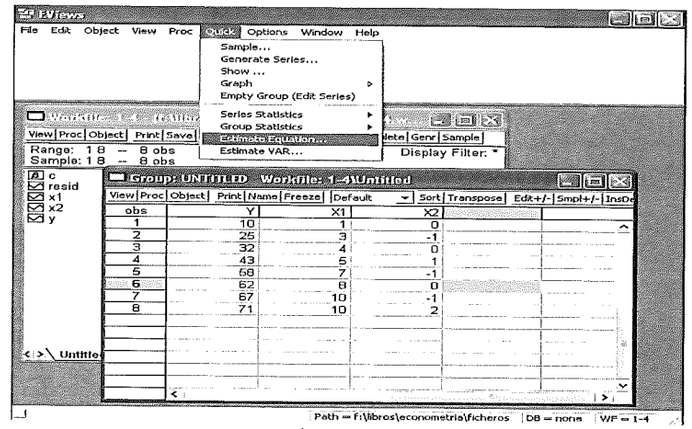


Figura 2-85

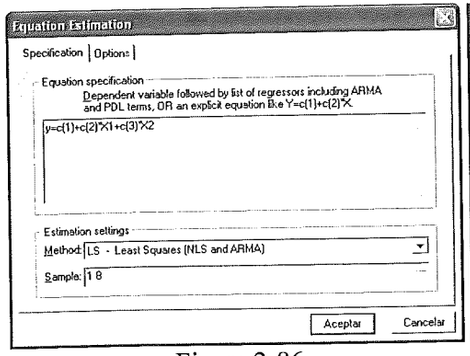


Figura 2-86

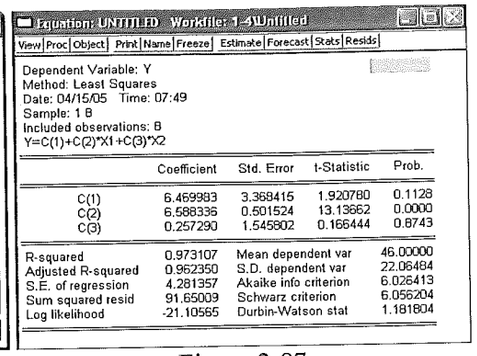


Figura 2-87

Para realizar los contrastes de restricciones sobre los coeficientes se utiliza *View→ Coefficient Tests→ Wald-Coefficient Restrictions* (Figura 2-88) y se rellenan las pantallas de entrada como se indica en las Figuras 2-89, 2-91, 2-93 y 2-95. El p-valor (*Probability*) pequeño de la figura 2-90 muestra el rechazo de la hipótesis $\beta_1=\beta_2=0$. El p-valor muy alto de la Figura 2-92 indica la aceptación de la hipótesis $\beta_1-10\beta_2=0$. El p-valor de la Figura 2-94 indica que se rechaza la hipótesis $2\beta_0+2\beta_1+7\beta_2=50$ al 85% de confianza. El p-valor de la Figura 2-96 muestra que se rechaza la hipótesis $\beta_1-10\beta_2=0$ y $2\beta_0+2\beta_1+7\beta_2=50$ simultáneamente al 95% de confianza.

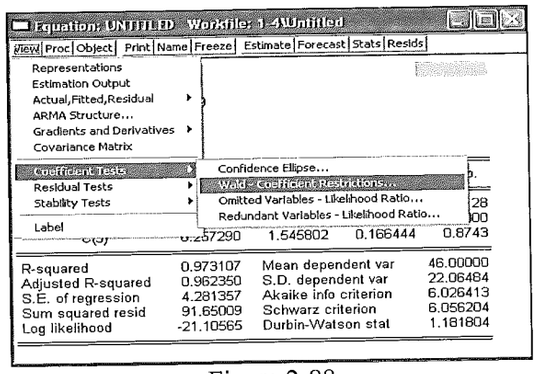


Figura 2-88

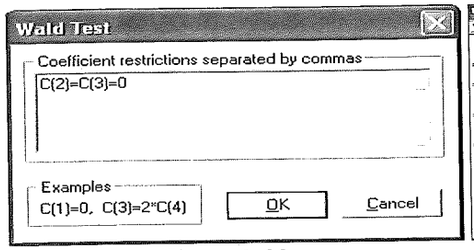


Figura 2-89

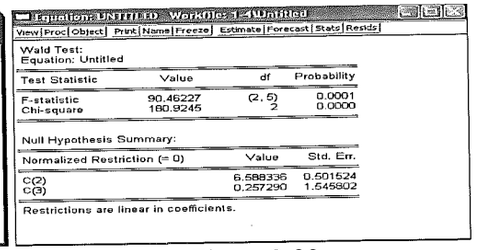


Figura 2-90

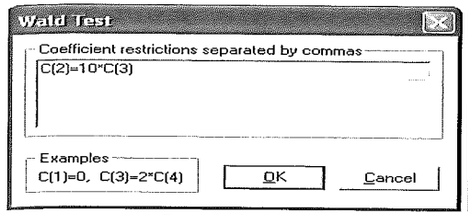


Figura 2-91

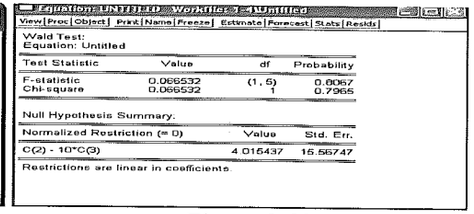


Figura 2-92

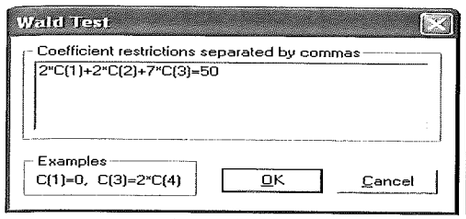


Figura 2-93

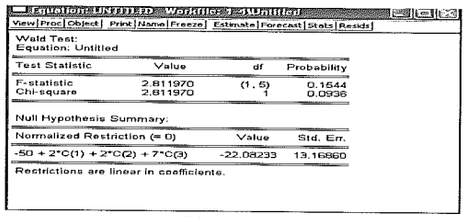


Figura 2-94

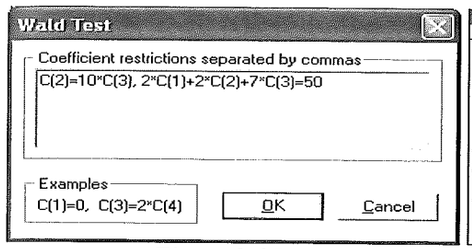


Figura 2-95

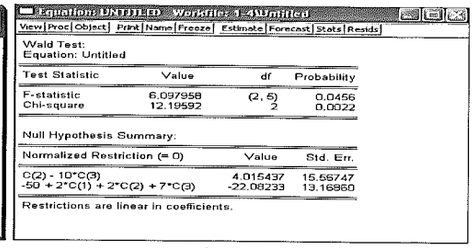


Figura 2-96

El contraste $\beta_0=\beta_1=\beta_2=0$ de significatividad conjunta de todos los parámetros del modelo puede obtenerse directamente si se rellena la pantalla *Equation Estimation* como se indica en la Figura 2-97 (en lugar de la Figura 2-86). Al pulsar *Aceptar* se obtiene la Figura 2-98, que presenta un p-valor para la *F* muy pequeño (*Pob(F-statistc)* = 0,000119), lo que nos lleva a aceptar la hipótesis de significatividad conjunta de los parámeros del modelo con probabilidad muy alta.

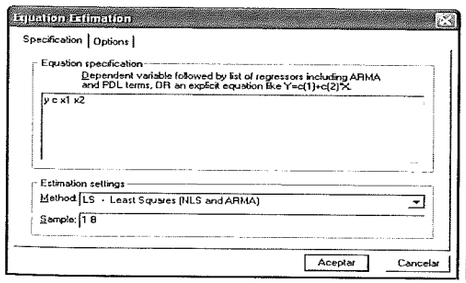


Figura 2-97

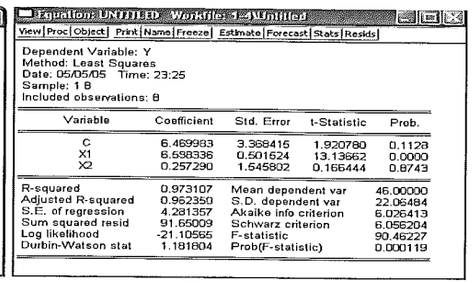


Figura 2-98

Realizamos ahora el ejercicio con el software STATA a partir del conjunto de datos *2-2.dta*. En primer lugar ajustamos la regresión de *y* sobre *x*1 y *x*2 obteniendo los resultados previamente estimamos con Eviews:

```
. use C:\LIBROS\2-2.dta

. regress y x1 x2

      Source |       SS       df       MS              Number of obs =       8
-------------+------------------------------           F(  2,     5) =   90.46
       Model |  3316.34991     2  1658.17496           Prob > F      =  0.0001
    Residual |  91.6500858     5  18.3300172           R-squared     =  0.9731
-------------+------------------------------           Adj R-squared =  0.9624
       Total |        3408     7  486.857143           Root MSE      =  4.2814

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   6.588336   .5015245    13.14   0.000     5.299126    7.877546
          x2 |   .2572899   1.545802     0.17   0.874    -3.716321    4.230901
       _cons |   6.469983   3.368415     1.92   0.113    -2.188803    15.12877
```

Para realizar los contrastes sobre las restricciones en los coeficientes utilizamos la sintaxis de STATA siguiente:

```
. test _cons=x1=x2=0

 ( 1)  - x1 + _cons = 0
 ( 2)  - x2 + _cons = 0
 ( 3)   _cons = 0

       F(  3,     5) =  368.15
            Prob > F =    0.0000

. test x1=x2=0

 ( 1)  x1 - x2 = 0
 ( 2)  x1 = 0

       F(  2,     5) =   90.46
            Prob > F =    0.0001

. test x1=10*x2

 ( 1)  x1 - 10 x2 = 0

       F(  1,     5) =    0.07
            Prob > F =    0.8067

. test 2*_cons+2*x1+7*x2=50

 ( 1)  2 x1 + 7 x2 + 2 _cons = 50

       F(  1,     5) =    2.81
            Prob > F =    0.1544
```

Se observa que se obtienen los mismos resultados que en Eviews.

Ejercicio 2-3. En un estudio para aumentar la supervivencia (variable y en tanto por ciento) de determinada clase de plantas se utilizan concentraciones de tres productos diferentes (variables x1, x2 y x3 en gramos). Para distintos valores de las variables x1, x2 y x3 se obtienen los siguientes porcentajes de aumento de la supervivencia:

y	x_1	x_2	x_3
25,5	1,74	5,30	10,80
31,2	6,32	5,42	9,40
25,9	6,22	8,41	7,20
38,4	10,52	4,63	8,50
18,4	1,19	11,60	9,40
26,7	1,22	5,85	9,90
26,4	4,10	6,62	8,00
25,9	6,32	8,72	9,10
32,0	4,08	4,42	8,70
25,2	4,15	7,60	9,20
39,7	10,15	4,83	9,40
35,7	1,72	3,12	7,60
26,5	1,70	5,30	8,20

Ajustar los datos a un modelo de regresión lineal múltiple que explique el porcentaje de aumento de la supervivencia en función de las concentraciones de los tres productos que inciden sobre la supervivencia de las plantas. Realizar la estimación de la varianza. Construir un intervalo de confianza al 95% para la respuesta media cuando x1 = 3, x2 = 8 y x3 = 9. Construir también un intervalo de predicción al 95% para una respuesta individual del porcentaje de supervivencia cuando x1 = 3, x2 = 8 y x3 = 9. Realizar el contraste de hipótesis b2 = -2,5 contra b2 > -2,5 para α = 0,05. Calcular el coeficiente de determinación R2 y realizar también el contraste de significación conjunta de la regresión y el contraste de significatividad individual parámetro a parámetro.

El trabajo relativo a este problema puede automatizarse con el software econométrico Eviews introduciendo los datos en un fichero de trabajo (2-3.*wk1*), cargándolo en el programa usando *File* → *Open* → *Eviews Workfile*...(Figura 2-99) y eligiendo el fichero en la ventana *Open* (Figura 2-100). Al pulsar *Abrir* se carga en memoria el fichero. Si queremos ver el contenido del conjunto de datos bastará con seleccionar todas las variables del modo habitual en Windows, hacer clic con el botón derecho del ratón sobre la selección y elegir *Open* → *As Group* (Figura 2-101). Se ve así el contenido de todas las variables del modelo como un grupo (Figura 2-102).

Para ***realizar la regresión***, se elige *Quick* → *Estimate Equation* (Figura 2-103), se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification* (Figura 2-104), se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar*. Se obtienen la Figura 2-105 con los resultados del modelo estimado.

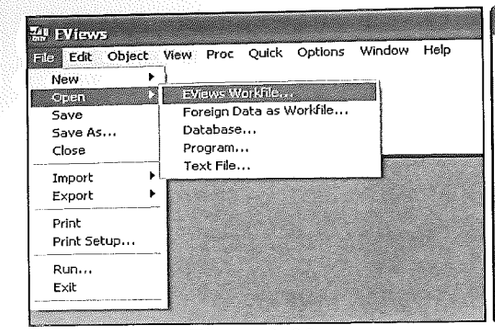


Figura 2-99

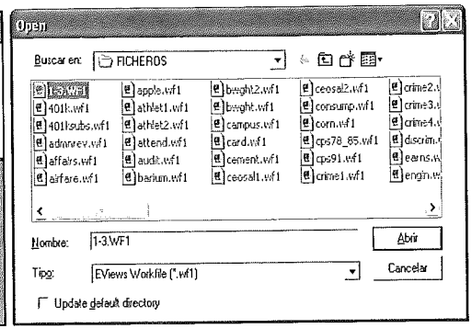


Figura 2-100

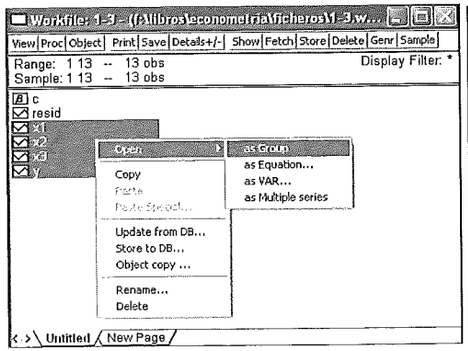


Figura 2-101

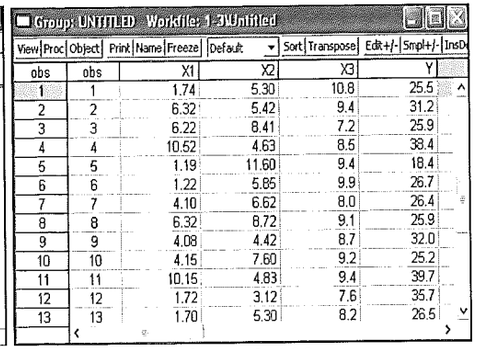


Figura 2-102

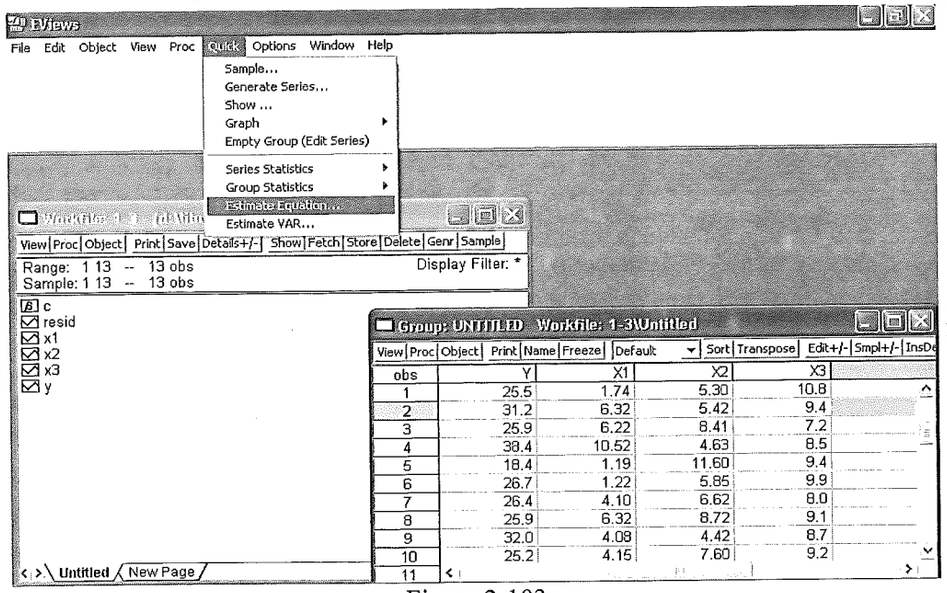


Figura 2-103

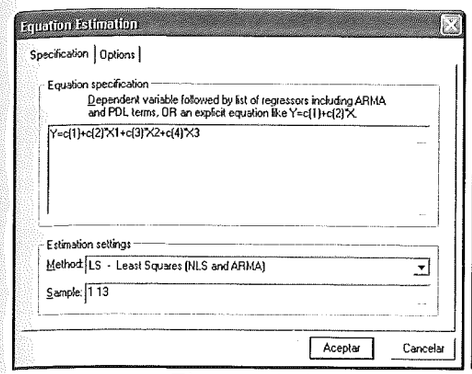


Figura 2-104

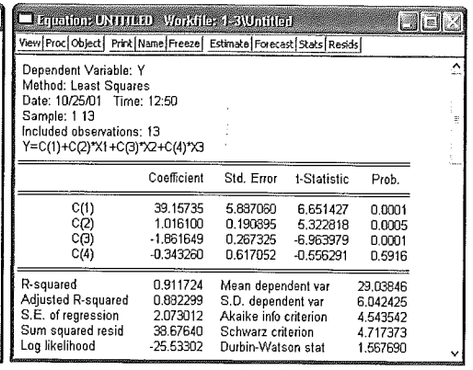


Figura 2-105

Observamos que las ***estimaciones de los parámetros*** (*Coefficient*), los valores del estadístico T (*t-Statistic*) y los p-valores (*Prob*) obtenidos a través de Eviews coinciden con los obtenidos realizando los cálculos directamente, pero la salida de Eviews es más rica. La columna *Std. Error* recoge la desviación típica estimada de los parámetros del modelo y puede utilizarse para el cálculo de ***intervalos de confianza de los parámetros estimados a un nivel*** α (*Coefficient* $\pm$ $t_{\alpha/2,T-k-1}$ *Std. Error*). Como el valor de la T de Student $t_{0,025} = 2{,}262$ para T-k-1 = 13-3-1 = 9 grados de libertad, tenemos que los intervalos de confianza al 95% para los parámetros estimados serán:

$$39{,}15735 \pm 2{,}262(5{,}887060)$$
$$1{,}016100 \pm 2{,}262(0{,}190895)$$
$$-1{,}86164 \pm 2{,}262(0{,}267325)$$
$$-0{,}34266 \pm 2{,}262(0{,}617052)$$

El R^2 ajustado vale 0,88 (indicativo de un buen ajuste al ser muy alto), el error estándar de la regresión (*S.E. of regresión*) es bajo y estima la desviación típica del error (su cuadrado es la varianza residual estimada $\hat{\sigma}^2 = 4{,}29$). Los valores bajos de los criterios de información de Akaike y Schwarz indican que el modelo es bueno. El valor del estadístico de Durbin Watson, no demasiado alejado de 2, indica que los ***problemas de autocorrelación*** no son relevantes. La suma de los errores al cuadrado (*Sum squared resid*) es el valor de la función objetivo en el mínimo cuando estimamos por mínimos cuadrados ordinarios. El logaritmo de la función de verosimilitud (*Log likelihood*) es el valor de la función objetivo en el máximo cuando se estima por máxima verosimilitud. Las características básicas de la variable dependiente vienen recogias por su media (*Mean dependent var*) y su cuasi desviación típica muestral (*S.D. dependent var*).

La ecuación del modelo definitivo ajustado es:

$$Y = 39{,}15 + 1{,}016X_1 - 1{,}86X_2 - 0{,}34\,X_3$$

La opción *View* → *Representations* (Figura 2-106) ofrece la ecuación del modelo estimado (Figura 2-107).

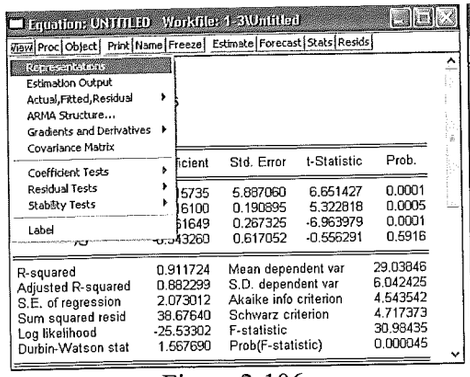


Figura 2-106

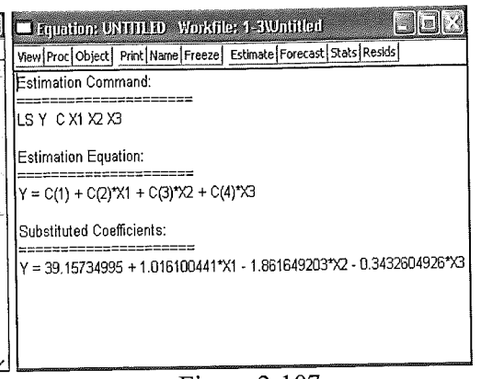


Figura 2-107

Al interpretar esta salida vemos que un aumento unitario en la concentración del primer producto produce un aumento de 1,016 unidades de la variable *Y*, es decir, un aumento de un 1,016% en la variable *Y* por venir esta última dada en tanto por ciento. De igual forma, un aumento unitario en la concentración del segundo producto produce un descenso de 1,86 unidades de la variable *Y*, es decir, un descenso de un 1,86% en la variable *Y*. Por último, un aumento unitario en la concentración del tercer producto produce un descenso del 0,34% en la variable *Y*.

En cuanto a la constante, podría interpretarse diciendo que para valores nulos de las concentraciones de los tres productos, el porcentaje de supervivencia de las plantas es del 39,15%, es decir, las plantas sobreviven casi con un 40% de probabilidad aunque no se le apliquen determinadas concentraciones de los tres productos. La constante juega aquí el papel de una cota fija de supervivencia de las plantas independiente de la aplicación o no de las correspondientes concentraciones de los tres productos.

El software Eviews permite calcular directamente la ***significatividad conjunta de los parámetros estimados del modelo*** a través del p-valor del contraste de la *F* siempre y cuando se introduzca la ecuación que define el modelo (pantalla *Equation Estimation*) tal y como se indica en la Figura 2-108 (en lugar de lo indicado en la Figura 2-104). Al pulsar *Aceptar*, se obtiene la Figura 2-109. Se observa un p-valor de la F muy pequeño (0,000045), lo que indica la aceptación de la hipótesis de la significatividad conjunta de todos los parámetros del modelo (pueden considerarse todos los parámetros del modelo significativamente distintos de cero simultáneamente con una probabilidad muy alta).

La ***significatividad individual de cada parámetro*** se analiza a través de los p-valores de la columna *Prob* en la Figura 1-11 (o en la Figura 2-105). Se ve que todos los parámetros estimados son fuertemente significativos individualmente (distintos de cero) por tener p-valores muy pequeños, salvo el parámetro de la variable X_3, que presenta una baja significatividad del 41% (p-valor = 0,59). Este hecho podría llevarnos a eliminar esta variable del modelo.

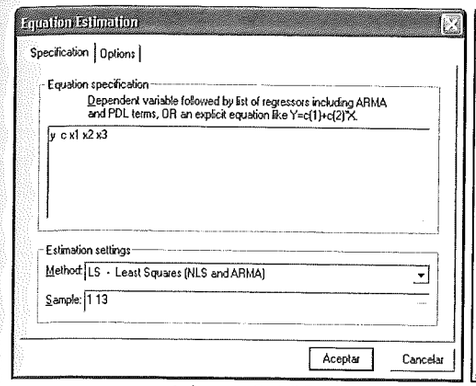


Figura 2-108

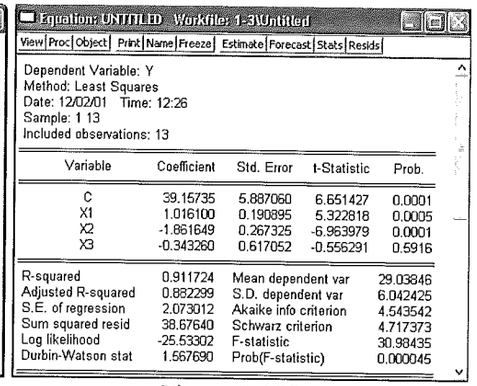


Figura 2-109

Eviews también permite ***contrastar restricciones en los parámetros*** a partir del Contraste de Wald mediante *View* → *Coefficient Tests*→ *Wald-Coefficient Restrictions* (Figura 2-110). Rellenando la pantalla *Wald Test* como se indica en la Figura 2-111 se contrasta la significatividad conjunta de los parámetros estimados por esta vía (alternativa al test de la *F*). En la Figura 2-112 se observa el rechazo de la nulidad simultánea de los parámetros porque el p-valor es muy pequeño, es decir, se acepta la significatividad conjunta de los parámetros del modelo. Por otra parte, si rellenamos la pantalla *Wald test* como se indica en la Figura 2-113, al pulsar *Aceptar* se obtiene la Figura 2-114, cuyo p-valor pequeño indica el rechazo de la hipótesis $b_2 = -2{,}5$.

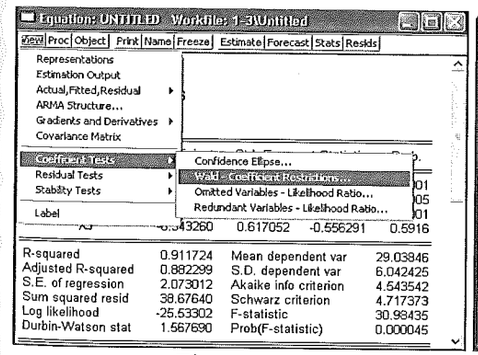


Figura 2-110

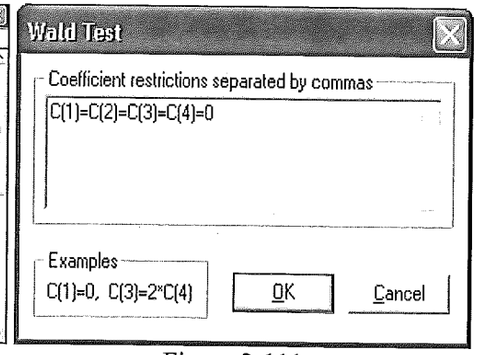


Figura 2-111

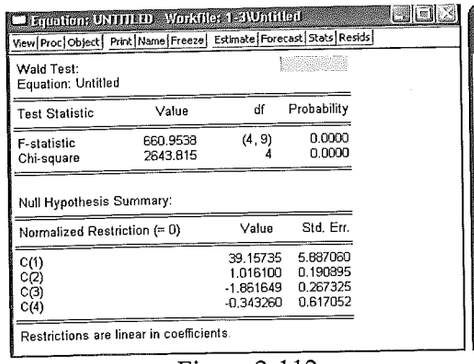


Figura 2-112

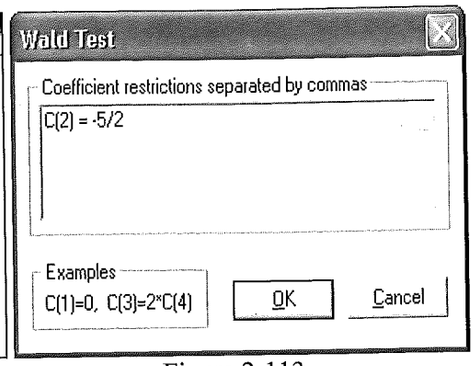


Figura 2-113

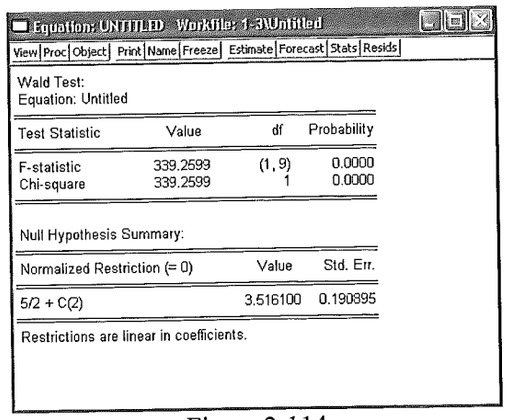


Figura 2-114

Para ***calcular predicciones*** (una en nuestro caso), nos situamos en la pantalla *Workfile* y elegimos *Proc → Structure/Resize Current Page* (Figura 2-115) para ampliar el tamaño de la muestra hasta el valor 14 (Figura 2-116). Al pulsar *OK* se obtiene el aviso de la Figura 2-117 indicativo de la inserción de una nueva observación en la muestra (que de momento no tiene valores). Si es necesario, con *Proc → Set Sample...* se amplía el tamaño de la muestra a 14. La siguiente tarea es introducir los valores de X_1, X_2 y X_3 para los que se obtendrá la predicción. Para ello seleccionamos como grupo las tres variables eligiendo *Open → as Group* en el menú emergente obtenido al hacer clic con el botón derecho del ratón sobe las tres variables seleccionadas simultáneamente (Figura 2-118). A continuación, en la pantalla *Group* se pulsa en *Edit* y se introducen los tres valores ee las variables (Figura 2-119). La tarea siguiente es elegir *Proc → Forecast* en la pantalla *Equation* (Figura 2-120) y rellenar la pantalla *Forecast* como se indica en la Figura 2-121 (El vector *Yf* contiene las predicciones y *SEf* sus errores estándar). Al pulsar *OK* aparecen los dos vectores en la pantalla *Workfile* (Figura 2-122).

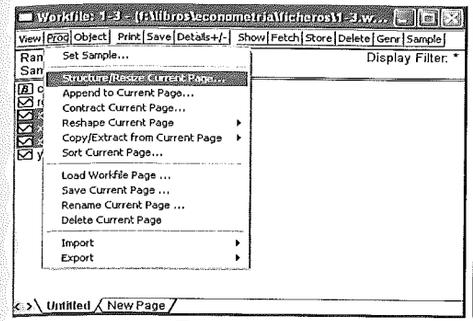


Figura 2-115

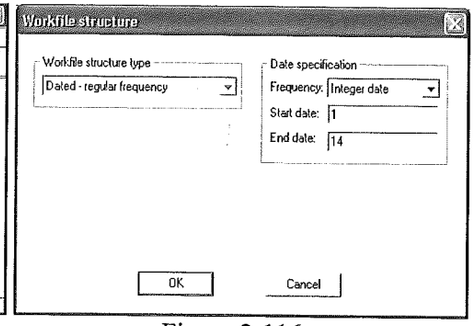


Figura 2-116

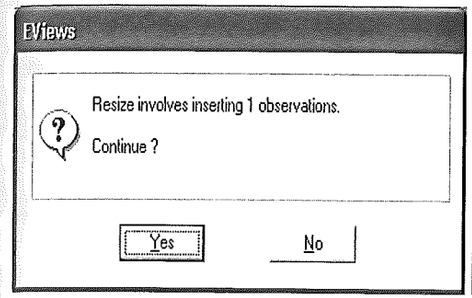


Figura 2-117

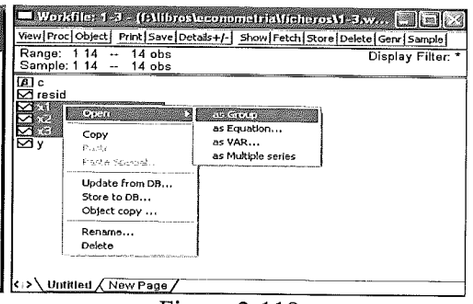


Figura 2-118

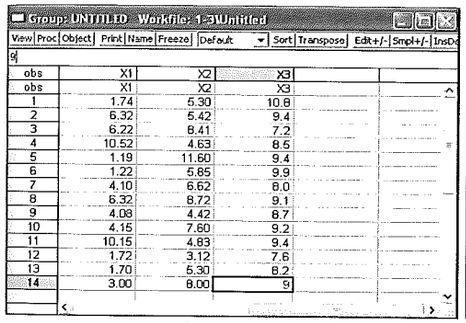


Figura 2-119

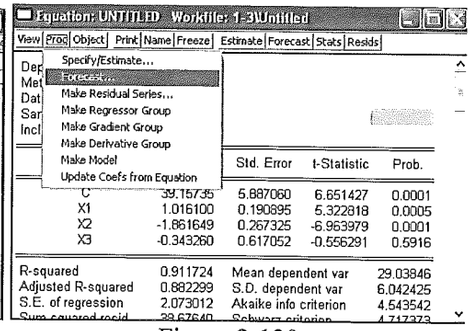


Figura 2-120

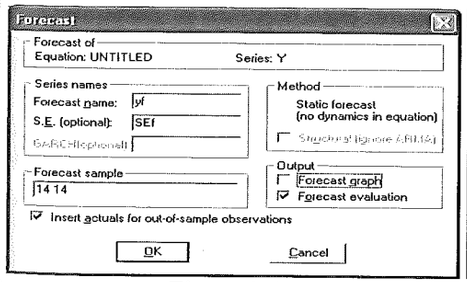


Figura 2-121

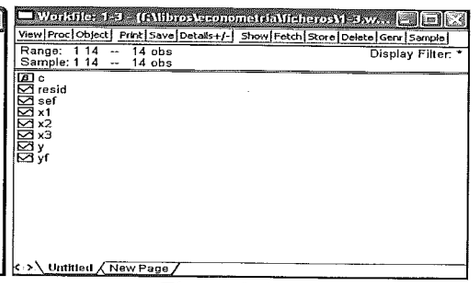


Figura 2-122

Al hacer clic sobre *yf* en la Figura 2-122 se obtiene la Figura 2-123, cuyo último valor es la predicción de la variable dependiente pedida (24,22311). Si se hace clic sobre *sef*, se obtiene la Figura 2-124, cuyo último valor es el error estándar de la predicción (2,209586), que se utilizará para calcular el intervalo de confianza para la predicción individual (Eviews no calcula el intervalo de confianza para la predicción en media).

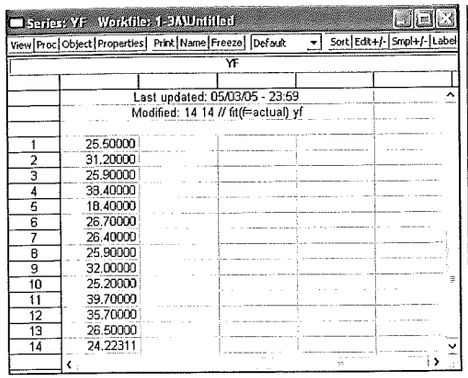


Figura 2-123

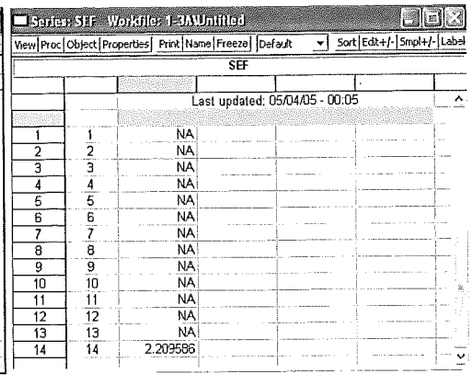


Figura 2-124

Como $t_{0,025}$ = 2,262 para T-k-1 = 13-3-1 = 9 grados de libertad, un ***intervalo de confianza de predicción para una respuesta individual*** cuando $x_1 = 3$, $x_2 = 8$ y $x_3 = 9$ vendrá dado por:

$$\hat{y}_0 \pm t_{\alpha/2}\underbrace{\hat{\sigma}\sqrt{1+x_0'(X'X)^{-1}x_0}}_{2,209586} = 24,22311 \pm (2,262)(2,209586) = [19,2 \quad 29,2]$$

Una vez estimado un modelo es muy convenirte realizar un ***análisis de los residuos*** para detectar posibles anomalías. Para ello, en la pantalla *Equation*, se utilizan las opciones de *Actual, Fitted, Residual* del menú *View* (Figura 2-125). La subopción *Actual, Fitted, Residual Table* ofrece los resultados de la Figura 2-126. Se obtienen los valores de la variable endógena observados Y_t y estimados $\hat{Y}_t$ y el valor de los residuos. El gráfico de los residuos presenta dos líneas discontinuas situados a una distancia de ±2S de la línea central continua, siendo S el error estándar de la regresión (estimador de la desviación típica del error). Estas bandas permiten detectar la presencia de residuos significativamente distintos de cero (atípicos). En nuestro caso los residuos números 5, 11 y 12 se salen de las bandas (aunque no demasiado), lo que puede indicar que son atípicos. La subopción *Actual, Fitted, Residual Graph* muestra el gráfico de la Figura 2-127 que representa los residuos en el eje de ordenadas de la izquierda y los valores de la variable endógena observados y estimados en el eje de ordenadas de la derecha. De esta forma, en la parte superior del gráfico aparecen las gráficas de los valores de la variable endógena observados y estimados, que han de coincidir lo más posible.

En la parte inferior aparece el gráfico de los residuos con las bandas ±2S, que coincide con el gráfico de la Figura 2-126. Este gráfico ha de ser lo más aleatorio posible sin comportamiento sistemático de ningún tipo para evitar la existencia de algún error de especificación en el modelo o la presencia de otros problemas como la autocorrelación. La opción *Residual Graph* de la Figura 2-125 presenta solamente el gráfico de los residuos y la opción *Standardized Residual Graph* presenta el gráfico de los residuos estandarizados (Figura 2-128). No se observan problemas de comportamiento sistemático residual por lo que puede suponerse aleatoriedad de los residuos. Los posibles valores atípicos pueden venir originados por la presencia de una variable poco significativa en el modelo (X_3).

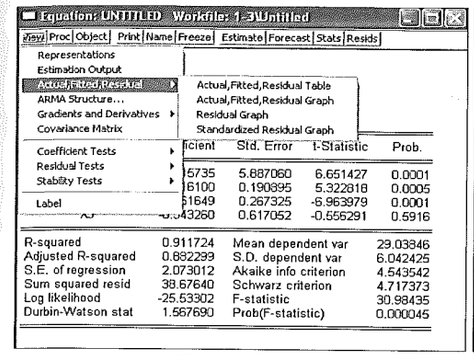


Figura 1-125

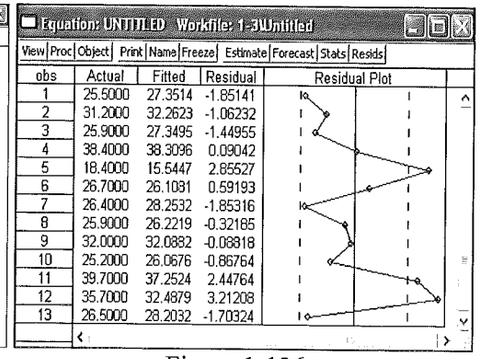


Figura 1-126

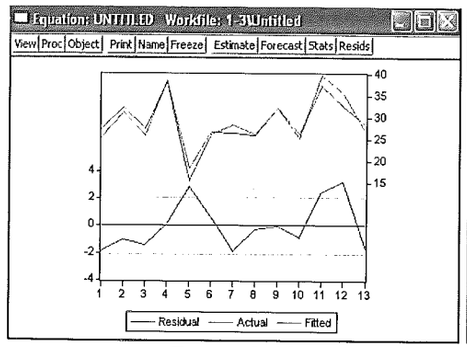


Figura 2-127

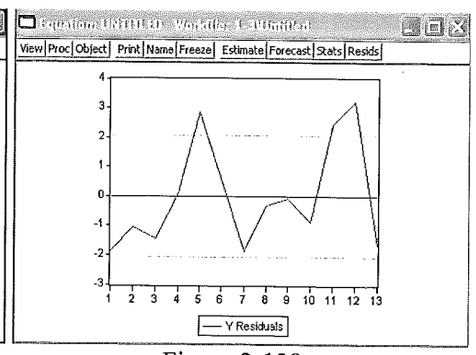


Figura 2-128

Ejercicio 2-4. Consideramos el fichero en formato STATA auto.dta con datos sobre automóviles que, entre otras, contiene las variables mpg (consumo), weight (peso) y price (precio). Se trata de estimar el modelo:

$$mpg = \beta_0 + \beta_1 price + \beta_2 weight + u$$

Contrastar la restricción paramétrica $\beta_1 = \beta_2$. Si la restricción es aceptable, estimar el modelo sujeto a esa restricción.

Realizaremos el ejercicio mediante la herramienta STATA.

```
. use auto
(1978 Automobile Data)

. regress mpg price weight

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  2,    71) =   66.85
       Model |  1595.93249     2  797.966246           Prob > F      =  0.0000
    Residual |  847.526967    71  11.9369995           R-squared     =  0.6531
-------------+------------------------------           Adj R-squared =  0.6434
       Total |  2443.45946    73  33.4720474           Root MSE      =   3.455

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -.0000935   .0001627    -0.57   0.567     -.000418    .0002309
      weight |  -.0058175   .0006175    -9.42   0.000    -.0070489   -.0045862
       _cons |   39.43966   1.621563    24.32   0.000     36.20635    42.67296
------------------------------------------------------------------------------

. test price=weight

 ( 1)  price - weight = 0

       F(  1,    71) =   63.49
            Prob > F =    0.0000
```

El p-valor del contraste permite aceptar la hipótesis de igualdad de parámetros contrastada. Por tanto es coherente realizar la estimación del modelo sujeto a dicha restricción. Para ello se utilizará el comando *cnsreg* mediante la siguiente sintaxis:

```
. constraint define 1 price = weight

. cnsreg mpg price weight, constraint(1)

Constrained linear regression                           Number of obs =      74
                                                        F(  1,    72) =   37.59
                                                        Prob > F      =  0.0000
                                                        Root MSE      =   4.722

 ( 1)  price - weight = 0
------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -.0009875   .0001611    -6.13   0.000    -.0013086   -.0006664
      weight |  -.0009875   .0001611    -6.13   0.000    -.0013086   -.0006664
       _cons |   30.36718   1.577958    19.24   0.000     27.22158    33.51278
```

Se observa que si $\beta_1 = \beta_2$ el modelo $mpg = \beta_0 + \beta_1 price + \beta_2 weight + u$ es equivalente a:

$$mpg = \beta_0 + \beta_1 price + \beta_1 weight + u = \beta_0 + \beta_1 (price + weight) + u$$

El modelo anterior se estima en STATA de la siguiente forma mediante el comando *regress*:


```
. generate x=price+weight

. regress mpg x

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  1,    72) =   37.59
       Model |  838.065767     1  838.065767           Prob > F      =  0.0000
    Residual |  1605.39369    72  22.2971346           R-squared     =  0.3430
-------------+------------------------------           Adj R-squared =  0.3339
       Total |  2443.45946    73  33.4720474           Root MSE      =   4.722

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.0009875   .0001611    -6.13   0.000    -.0013086   -.0006664
       _cons |   30.36718   1.577958    19.24   0.000     27.22158    33.51278
```

Se observa que el resultado es el mismo que en la regresión con restricciones.

Ejercicio 2-5. El archivo 2-5.wf1 contiene datos trimestrales correspondientes a los ejercicios 1996-2003, relativos al consumo de electricidad en España Yt medido en GWh y al PIB a precios de mercado en millones de euros constantes de 1995.

Se trata de ajustar un modelo adecuado que explique el consumo de electricidad en función del PIB teniendo presente la posible estacionalidad de los datos.

Existe la posibilidad de que el consumo de energía eléctrica Y_t sea una variable estacional trimestral. Para representarla gráficamente utilizamos *Quick* → *Graph* → *Scatter*, rellenamos la pantalla *Series List* como se indica en la Figura 2-129 y al pulsar *OK* se obtiene el gráfico de la Figura 2-130, cuya estructura indica la estacionalidad trimestral.

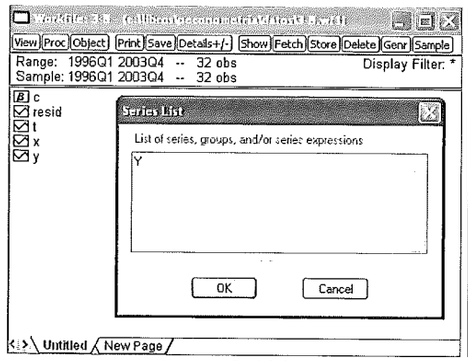


Figura 2-129

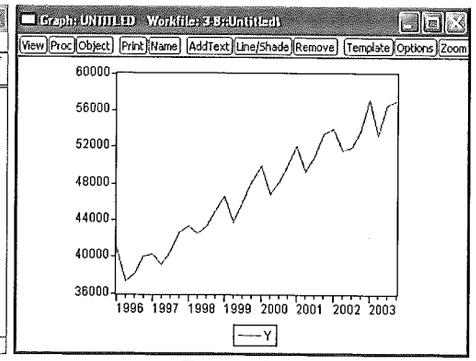


Figura 2-130

Dado que la serie temporal Y_t es estacional trimestral, para tener presentes los efectos de la estacionalidad consideramos el modelo:

$$Y_t = \alpha_0 + \alpha_1 Q_{1t} + \alpha_2 Q_{2t} + \alpha_3 Q_{3t} + \alpha_4 X_t + u_t$$

$$Q_{it} = \begin{cases} 1 \; en \; el \; trimestre \; i \\ 0 \; en \; el \; resto \end{cases} \qquad i = 1, \cdots, 4$$

Se observa que en el modelo se omite la variable ficticia relativa al cuarto trimestre para evitar la colinealidad perfecta.

Ajustamos el modelo por MCO rellenando la pantalla *Equation Estimation* de Eviews como se indica en la Figura 2-131 obteniendo los resultados de la Figura 2-132.

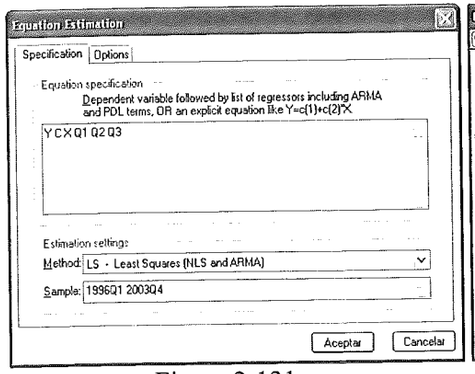


Figura 2-131

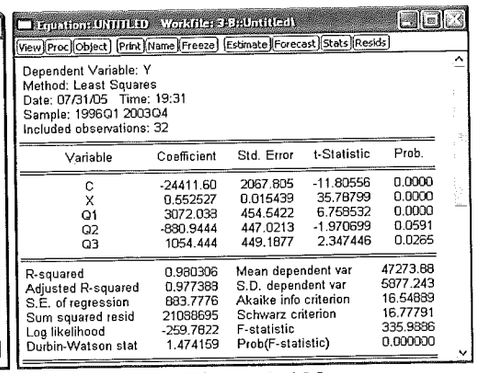


Figura 2-132

Se observa significatividad individual y conjunta de los parámetros muy alta y muy buen coeficiente de determinación. El estadístico de Durbin-Watson presenta un valor no muy alejado de 2, lo que puede indicar ausencia de autocorrelación.

Como los parámetros estimados del modelo anterior relativos a las variables ficticias son significativamente distintos de cero individualmente, el efecto estacional está presente. Estos parámetros miden el efecto relativo de cada trimestre. El modelo finalmente estimado es el siguiente:

$$Y_t = -24411.59789 + 3072.03843\, Q_{1t} - 880.9443838\, Q_{2t} + 1054.44384\, Q_{3t} + 0.552526998\, X_t + u_t$$

Ejercicio 2-6. Supongamos que el coste de las comisiones Y pagadas por una cierta empresa varía proporcionalmente con las ventas X del producto. Se sabe que a partir de un cierto nivel de éstas, por ejemplo 5500, la proporción de la comisión se eleva. Se dispone de los datos siguientes recogidos en el archivo 2-6.wf1.

Y	256	414	634	778	1003	1839	2081	2423	2734	2914
X	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000

Estimar un modelo de regresión adecuado que explique las comisiones pagadas en función de las ventas del producto.

Como a partir de ventas superiores a 5500 la comisión se eleva, la pendiente del modelo ajustado será diferente antes y después de esa cantidad. Por esta razón utilizaremos una ***regresión por tramos***. Considerar dos tramos de la variable X (X<5500 y X>5500).

Para realizar la estimación del modelo se creará una variable ficticia D que valga cero para valores inferiores a 5500 y 1 para los valores superiores.

La ecuación del modelo a ajustar será la siguiente:

$$Y_t = \alpha_1 + \alpha_2 X_t + \alpha_3 (X_t - 5500) D_t + u_t$$

$$D_t = \begin{cases} 1 \; si \; X_t > 5500 \\ 0 \; en \; el \; resto \end{cases}$$

Suponiendo E(u_t)=0, se tiene que:

$$E(Y_t \,|D_t = 0, X_t, 5500) = \alpha_1 + \alpha_2 X_t$$
$$E(Y_t \,|D_t = 1, X_t, 5500) = \alpha_1 - \alpha_3 X^* + (\alpha_2 + \alpha_3) X_t$$

Se observa que α_2 corresponde a la pendiente de la recta de regresión en el primer tramo (X<5500) y $\alpha_2+\alpha_3$ es la pendiente de la recta de regresión en el segundo tramo (X>5500).

Ajustamos el modelo por MCO rellenando la pantalla *Equation Estimation* de Eviews como se indica en la Figura 2-133, se obtienen los resultados de la Figura 2-134. La variable (X-5500)*D se ha denominado *TR*. Se observa mala significatividad individual para *C* y *TR*, significatividad conjunta muy alta y muy buen coeficiente de determinación.

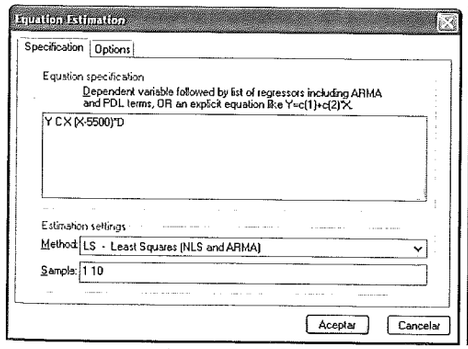


Figura 2-133

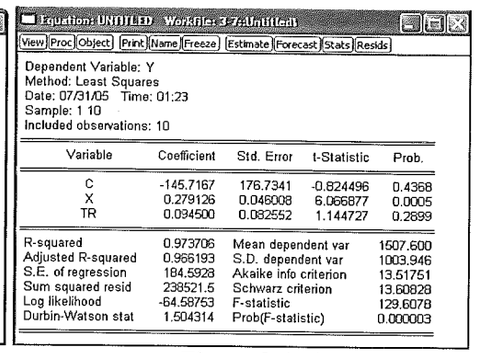


Figura 2-134

Los modelos ajustados en ambos tramos son los siguientes:

$Y_t = -145{,}7167 + 0{,}279126\,X_t$ si $X_t < 5500$
$Y_t = -145{,}7167 - 0{,}0945*5500 + (0{,}279126 + 0{,}0945)X_t = 665{,}466 + 0{,}3736\,X_t$ si $X_t > 5500$

Podrían ajustarse los modelos sin constante por MCO rellenando la pantalla *Equation Estimation* de Eviews como se indica en la Figura 2-135 para obtener los resultados de la Figura 2-136. Ahora todos los parámetros estimados son significativos.

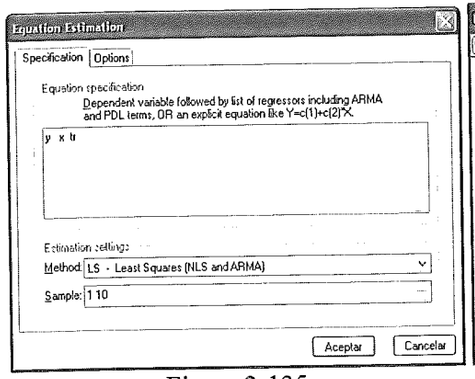


Figura 2-135

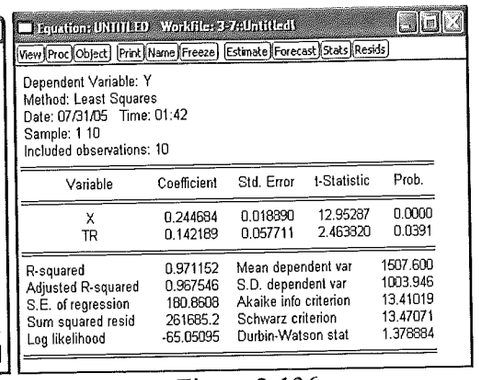


Figura 2-136

Los modelos ajustados en ambos tramos son los siguientes:

$Y_t = 0{,}244684\,X_t$ si $X_t < 5500$
$Y_t = 0{,}142189*5500 + (0{,}244684 + 0{,}142189)X_t = 782{,}04 + 0{,}386873\,X_t$ si $X_t > 5500$

Ejercicio 2-7. Con el objetivo de estudiar la evolución fiscal en España se estima un modelo de regresión que intenta explicar las variaciones de los ingresos públicos a partir del PIB nominal. Se dispone de la siguiente información del Banco de España referida al periodo 1955-1991:

Año	*Ingresos*	*PIB*	*Año*	*Ingresos*	*PIB*
1960	*72651*	*681900*	*1976*	*844320*	*7266390*
1961	*81067*	*777510*	*1977*	*1 135047*	*9219920*
1962	*94950*	*898130*	*1978*	*1 407981*	*11285000*
1963	*109108*	*1062340*	*1979*	*1 762400*	*13201100*
1964	*125844*	*1208820*	*1980*	*2 179100*	*15167970*
1965	*149538*	*1402220*	*1981*	*2 574600*	*17044800*
1966	*185282*	*1626690*	*1982*	*2 975600*	*19722640*
1967	*213348*	*1842050*	*1983*	*3 618300*	*22531770*
1968	*239237*	*2079630*	*1984*	*4 218400*	*25519540*
1969	*270542*	*2381190*	*1985*	*4 775600*	*28200890*
1970	*303999*	*2629860*	*1986*	*6 039300*	*32323990*
1971	*364034*	*2967990*	*1987*	*7 115500*	*36143970*
1972	*418709*	*3483420*	*1988*	*8 185000*	*40163970*
1973	*483975*	*4199410*	*1989*	*9 847400*	*45024900*
1974	*570538*	*5143300*	*1990*	*10 435700*	*50074120*
1975	*682646*	*6052640*	*1991*	*11 481600*	*54775380*

Con las variables en logaritmos, estime un modelo de regresión simple entre los ingresos públicos (Ingresos) y el PIB nominal (PIB). Incluir en la estimación una variable ficticia, de modo aditivo y multiplicativo, que toma el valor 1 desde el año 1978 en adelante. Realizar la misma estimación valorándola estadísticamente. Representar de nuevo el gráfico de residuos comparándolo con el del apartado anterior.

Comenzamos introduciendo los datos del problema en dos variables llamadas INGRESOS y PIBN del fichero de datos en formato SPSS de nombre 2_7.sav. También introducimos en el fichero los logaritmos de las variables anteriores (LOG_INGRESOS y LOG_PIBN) y la variable FICTICIA1 que vale 0, entre 1960 y 1977, y 1 entre 1978 y 1991 (variable que recoge el cambio estructural consecuencia del impacto de la crisis energética de 1977 en la economía española). También se introduce la variable producto de FICTICIA1 por logaritmo del PIB (FICTICIA1_LOG_PIBN).

Para ajustar el modelo de regresión que explica el logaritmo del PIB en función del logaritmos de los ingresos con SPSS, utilizamos la ruta *Analizar* → *Regresión* → *Lineal* (Figura 2-137) y rellenamos la pantalla de entrada como se indica en la Figura 2-138. Las pantallas correspondientes a los botones *Estadísticos, Gráficos* y *Opciones* se rellenan como se indica en las Figuras 2-139 a 2-141 Al pulsar Aceptar se obtienen los resultados de la regresión de la Figura 2-142.

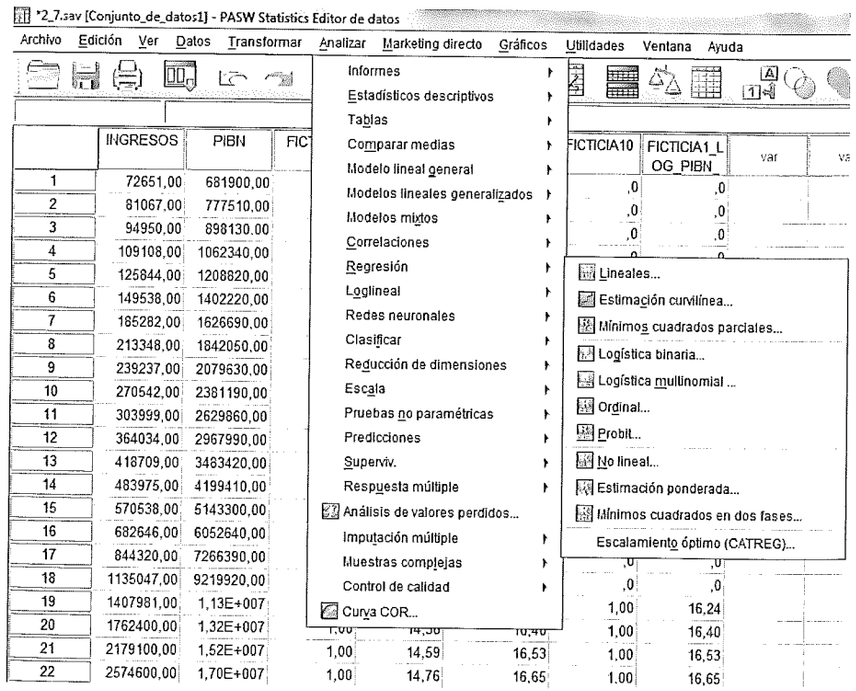


Figura 2-137

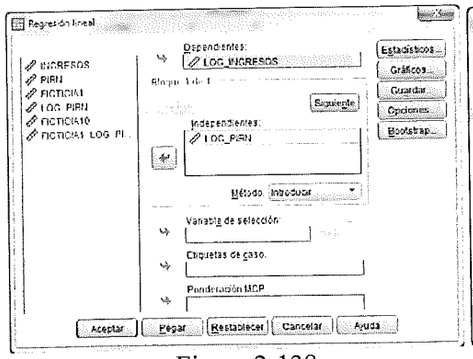


Figura 2-138

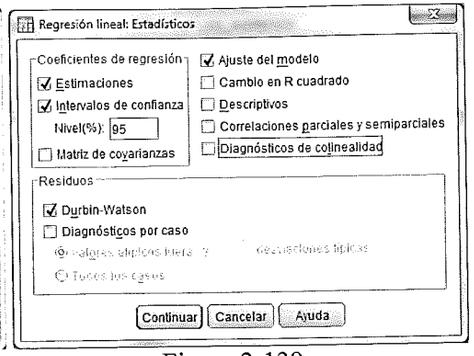


Figura 2-139

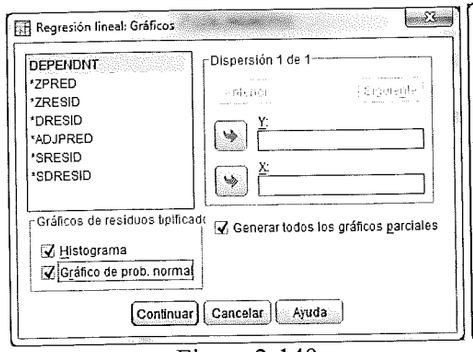


Figura 2-140

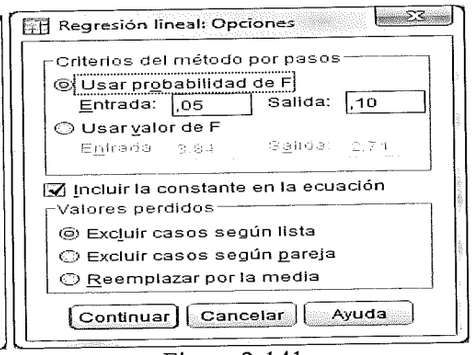


Figura 2-141

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,998[a]	,997	,996	,09698018	,162

a. Variables predictoras: (Constante), LOG_PIBN

b. Variable dependiente: LOG_INGRESOS

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	81,589	1	81,589	8674,973	,000[a]
	Residual	,282	30	,009		
	Total	81,872	31			

a. Variables predictoras: (Constante), LOG_PIBN

b. Variable dependiente: LOG_INGRESOS

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	-4,534	,197		-23,062	,000	-4,935	-4,132
	LOG_PIBN	1,161	,012	,998	93,140	,000	1,136	1,187

a. Variable dependiente: LOG_INGRESOS

Figura 2-142

Según los coeficientes estimados de la Figura 2-142, la ecuación de ajuste será la siguiente:

$$LOG(INGRESOS) = -4{,}534 + 1{,}161\ LOG(PIBN)$$

La pendiente del modelo y la constante resultan significativos al 95% (se rechaza la hipótesis nula de que valgan cero), ya que los *p*-valores son menores que 0,05. Además, tanto el coeficiente de determinación como el coeficiente de correlación son muy altos, lo que favorece el ajuste. El contraste global del modelo también es muy bueno (valor de la *F* mayor que su valor crítico). El histograma de los residuos (Figura 2-143) se ajusta bien a la campana de Gauss y el gráfico de normalidad (Figura 2-144) se ajusta bastante bien a la diagonal del primer cuadrante. El modelo puede estar mal especificado.

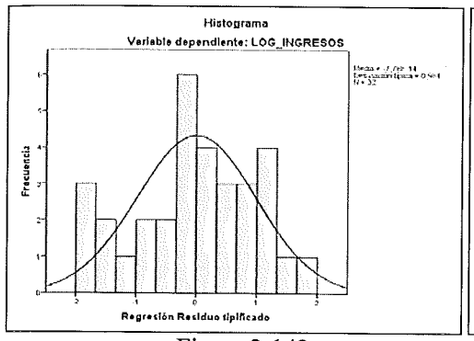


Figura 2-143

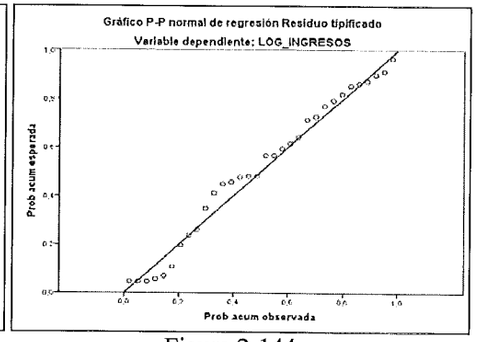


Figura 2-144

Pero el estadístico de Durbin-Watson que se observa en el apartado *Resumen del modelo* de la Figura 2-142 toma el valor 0,162 muy alejado del valor ideal 2. Lo que indica presencia de autocorrelación acusada. Más concretamente, pata n=27, k=4 y α=0,05, el valor de los los estadísticos superior e inferior de la tabla D-W son d_L=1,08 y d_U=1,7. Como 0,162 es mucho menor que d_L=1,08, existe autocorrelación acusada.

Para intentar resolver el problema de autocorrelación, introducimos la variable de nombre FICTICIA1, que vale 0, entre 1960 y 1977, y 1 entre 1978 y 1991 (variable que recoge el cambio estructural consecuencia del impacto de la crisis energética de 1977 en la economía española). Dicha variable la introducimos de forma aditiva y multiplicativa, dando lugar al modelo siguiente:

$$LOG(INGRESOS)=\beta_1+\beta_2*FICTICIA1+\beta_3*LOG(PIBN)+\beta_4*FICTICIA1*LOG(PIBN)+\mu$$

El nuevo modelo propuesto se estima rellenado la pantalla de entrada del procedimiento *Regresión* de SPSS como se indica en la Figura 2-145. Al pulsar *Aceptar*, se obtienen los resultados numéricos de la Figura 2-147, que presenta un buen ajuste con p-valores pequeños para la significatividad individual y conjunta de los parámetros estimados y un estadístico de Durbin-Watson de valor 0,963 bastante mejorado (prácticamente alcanza el valor de d_L=1,08, lo que indica que se han arreglado los problemas de autocorrelación). También mejora R^2 y la distribución de los residuos (Figura 2-146).

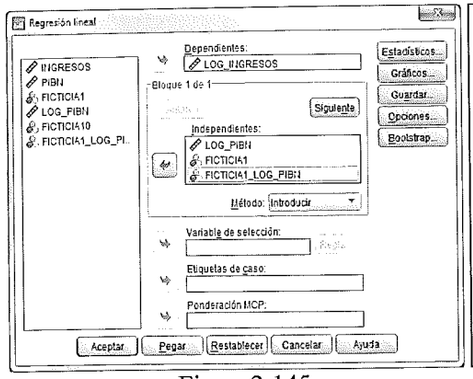


Figura 2-145

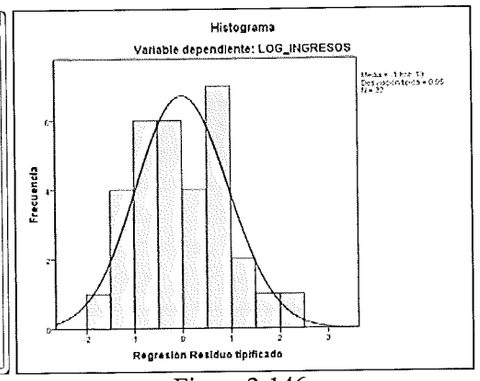


Figura 2-146

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	1,000[a]	1,000	1,000	,03598820	,963

a. Variables predictoras: (Constante), FICTICIA1_LOG_PIBN, LOG_PIBN, FICTICIA1

b. Variable dependiente: LOG_INGRESOS

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	81,835	3	27,278	21061,996	,000[a]
	Residual	,036	28	,001		
	Total	81,872	31			

a. Variables predictoras: (Constante), FICTICIA1_LOG_PIBN, LOG_PIBN, FICTICIA1

b. Variable dependiente: LOG_INGRESOS

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	-2,976	,163		-18,208	,000	-3,311	-2,641
	LOG_PIBN	1,054	,011	,906	94,564	,000	1,031	1,077
	FICTICIA1	-4,790	,373	-1,486	-12,834	,000	-5,554	-4,025
	FICTICIA1_LOG_PIBN	,297	,023	1,575	13,162	,000	,251	,343

a. Variable dependiente: LOG_INGRESOS

Figura 2-147

El nuevo modelo estimado presenta la ecuación siguiente:

$$\text{LOG(INGRESOS)} = -2{,}97 - 4{,}79*\text{FICTICIA1} + 1{,}054*\text{LOG(PIBN)} + 0{,}297*\text{FICTICIA1}*\text{LOG(PIBN)}$$

Está claro que este modelo proporciona un ajuste mejorado, explicando un 99,96 % de la variación total. La hipótesis de nulidad del conjunto de parámetros se rechaza claramente al 95% en el contraste de la F (p-valor menor que 0,05). Lo mismo ocurre con la hipótesis de nulidad de cada parámetro en particular, resultando todas las variables significativas (p-valores menores que 0,05).

Ejercicio 2-8. La producción de la minería española entre los años 1969 y 1984 expresada en unidades monetarias constantes de 1984 toma los valores Xt de la tabla adjunta. El empleo del factor trabajo en la producción se expresa mediante la variable Wt que cuantifica los millones de horas/hombre trabajadas. Para medir el stock de capital o riqueza se utiliza la variable Ct que representa la potencia instalada en miles de caballos de vapor.

X_t	W_t	C_t
179,2	193,5	1141
181,0	182,8	1241
183,1	171,7	1357
184,9	163,4	1465
185,8	143,3	1562
220,8	140,4	1742
238,8	141,6	1954
241,7	138,6	2141
242,5	145,4	2352
240,7	128,1	2399
248,5	126,4	2557
312,1	149,2	2680
347,3	145,9	2899
366,2	144,5	3082
424,7	139,7	3062
404,9	131,8	3052

Se trata de estimar las elasticidades del trabajo y el capital respecto de la producción de nuestra industria minera en el período 1964-84 considerando como modelo la función de producción de Cobb-Douglas.

El modelo de Cobb-Douglas será: $X_t = kW_t^a C_t^b$

Como se trata de un modelo no lineal, aplicamos logaritmos para transformarlo en el modelo lineal siguiente:

$$LogX_t = Logk + aLogW_t + bLogC_t + u_t$$

Para resolver este problema con Eviews, se carga el fichero de trabajo (2-8.*wk*1) usando *File* → *Open* → *Eviews Workfile...* y se elige *Quick* → *Estimate Equation* (Figura 2-148). Se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar* (Figura 2-149).

Se obtienen los resultados de la Figura 2-150. Se observa que todos los parámetros estimados (incluida la constante) son significativamente distintos de cero (p-valores muy pequeños que indican una significatividad del 99% para el parámetro de W_t, y superior para los demás). El R^2 y el R^2 ajustado son 0,91 y 0,9 respectivamente (muy altos).

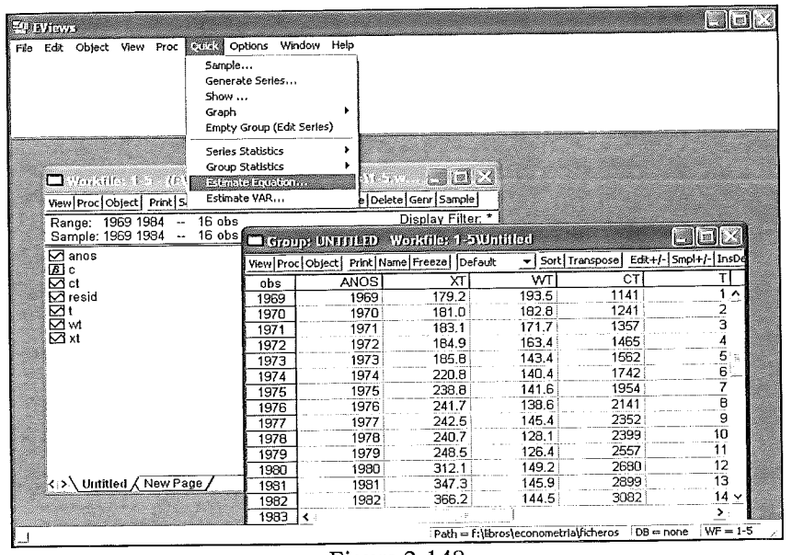


Figura 2-148

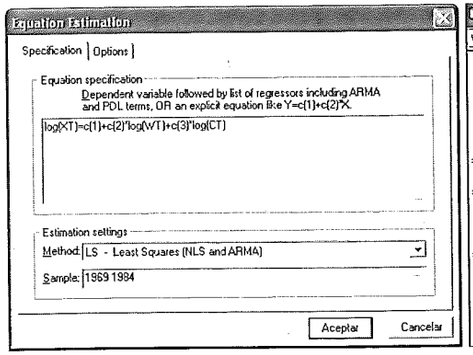


Figura 2-149

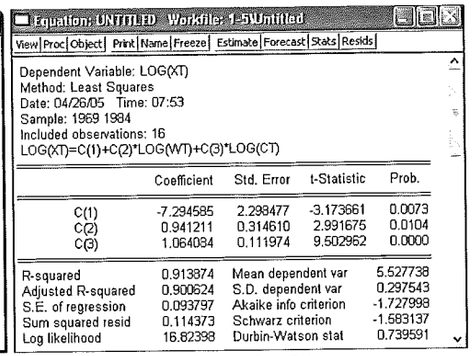


Figura 2-150

El modelo estimado es el siguiente:

$$LogX_t = -7{,}29+0{,}94LogW_t+1{,}06LogC_t$$

que puede escribirse de la forma:

$$X_t = 0{,}00068W_t^{\,0{,}94} + C_t^{\,1{,}06}$$

Como el modelo ha sido estimado con todas las variables en logaritmos, podemos interpretar los parámetros estimados en términos de elasticidades. Así, la elasticidad trabajo-producto se ha estimado en 0,94, lo que indica que si se incrementa el empleo de la minería en un 1%, la producción cae algo menos del 1% (0,94%). Además, la elasticidad capital-producto es algo superior a la unidad (1,06), lo que indica que incrementos del capital (potencia instalada) del 1% generan incrementos productivos superiores a ese porcentaje (1,06%).

Se observa que el R^2 y el R^2 ajustados son muy altos explicando las variables exógenas más de un 90% de la variabilidad de la variable endógena. La desviación típica estimada del error es 0,093797 y los criterios de información de Akaike y Schwartz tienen valores pequeños con lo que la capacidad explicativa del modelo es buena.

El estadístico de Durban Watson tiene un valor demasiado alejado de 2, lo que puede provocar posibles problemas de autocorrelación.La significatividad individual de los parámetros es buena porque los p-valores son menores que 0,05. La significatividad más pequeña es casi del 99% (variable *LogW*). Para ver la significatividad conjunta, rellenamos la pantalla *Equation Estimation* como se indica en la Figura 2-151. Al pulsar *Aceptar* se obtienen los resultados de la Figura 2-152 observándose un p-valor de la F muy bajo, lo que indica alta significatividad conjunta de los parámetros del modelo.

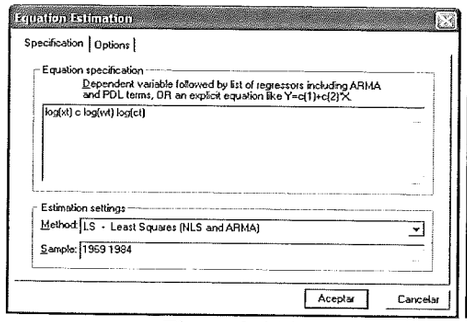


Figura 2-151

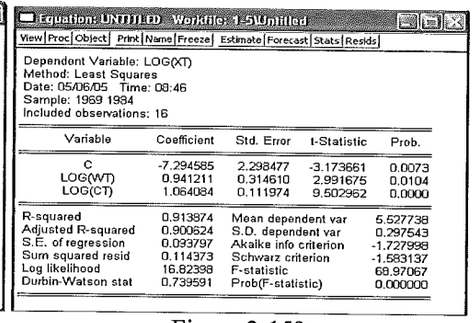


Figura 2-152

Para analizar la importancia de la posible autocorrelación graficamos los valores observados de la variable dependiente contra los residuos a partir de la opción *Quick* → *Graph* → *Scatter* (Figura 2-153). Se rellena la pantalla *Series list* con los residuos y la variable dependiente (Figura 2-154). Al pulsar *OK* se obtiene el gráfico de dispersión de la Figura 2-155, que presenta una estructura aleatoria de sus puntos indicando la ausencia de problemas serios de autocorrelación.

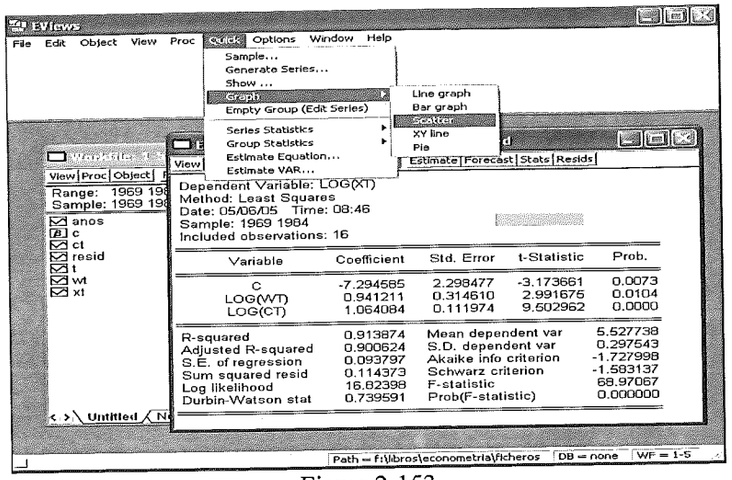


Figura 2-153

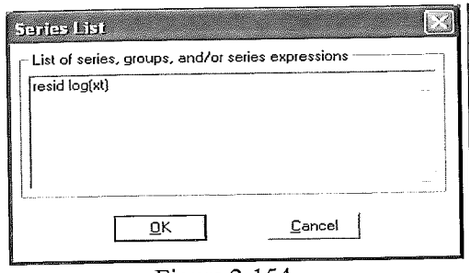


Figura 2-154

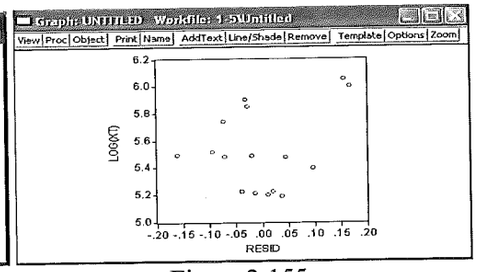


Figura 2-155

Ejercicio 2-9. Se intenta medir el grado de salud en un grupo de personas de diferentes edades controlando el consumo de oxígeno (oxigen) en unas pruebas físicas en función de su edad, peso, tiempo empleado en la realización de las pruebas, pulsaciones en descanso (pd), pulsaciones en ejercicio (pe) y pulsaciones máximas (pm). Para ello se quiere realizar un ajuste del conjunto de datos de que se dispone para estas variables (contenido en el fichero en formato SPSS salud.sav) a un modelo lineal que contenga las variables precisas, de modo que la calidad del ajuste sea razononable. Se utilizará el método automático de selección de variables hacia atrás (backward) para el modelo de regresión. Inicialmente se supone que el modelo podría tener la forma: oxigen=a+b*edad+c*peso+ d*tiempo+e*pd+f*pe+g*pm.

Utilizamos la ruta *Analizar* → *Regresión* → *Lineal* (Figura 2-137) y rellenamos la pantalla de entrada como se indica en la Figura 2-156. Observar que en el campo Método se ha elegido Atrás (regresión *backward*). Las pantallas correspondientes a los botones *Estadísticos y Gráficos* se rellenan como se indica en las Figuras 2-157 y 2-158. Al pulsar *Aceptar* se obtienen los resultados de la regresión (Figura 2-159 y 2-160).

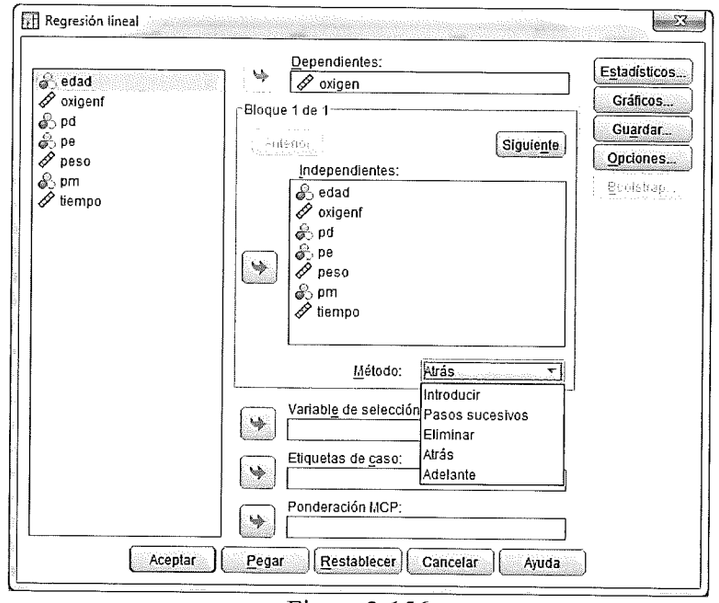


Figura 2-156

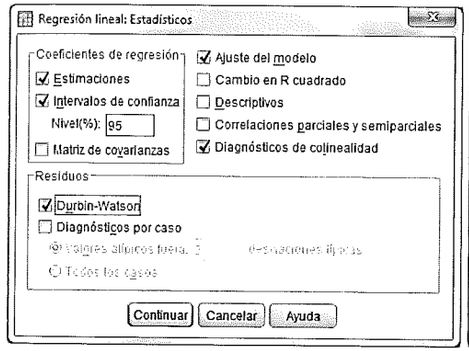


Figura 2-157

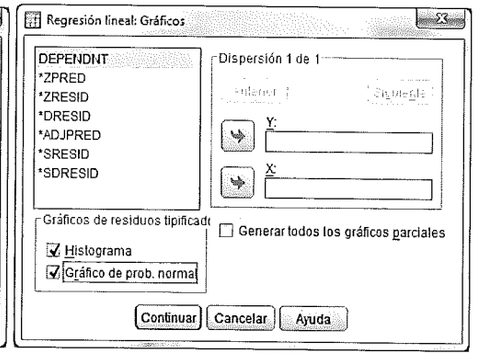


Figura 2-158

Variables introducidas/eliminadas[b]

Modelo	Variables introducidas	Variables eliminadas	Método
1	tiempo, peso, pm, pd, edad, pe[a]	.	Introducir
2	.	pd	Hacia atrás (criterio: Prob. de F para salir >= ,100).
3	.	peso	Hacia atrás (criterio: Prob. de F para salir >= ,100).

a. Alcanzado límite de tolerancia = ,000.

b. Variable dependiente: oxigen

Figura 2-159

Resumen del modelo[d]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,921[a]	,849	,811	2,31695	
2	,921[b]	,848	,818	2,27516	
3	,915[c]	,837	,812	2,31159	1,791

a. Variables predictoras: (Constante), tiempo, peso, pm, pd, edad, pe
b. Variables predictoras: (Constante), tiempo, peso, pm, edad, pe
c. Variables predictoras: (Constante), tiempo, pm, edad, pe
d. Variable dependiente: oxigen

ANOVA[d]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	722,544	6	120,424	22,433	,000[a]
	Residual	128,838	24	5,368		
	Total	851,382	30			
2	Regresión	721,973	5	144,395	27,895	,000[b]
	Residual	129,408	25	5,176		
	Total	851,382	30			
3	Regresión	712,452	4	178,113	33,333	,000[c]
	Residual	138,930	26	5,343		
	Total	851,382	30			

a. Variables predictoras: (Constante), tiempo, peso, pm, pd, edad, pe
b. Variables predictoras: (Constante), tiempo, peso, pm, edad, pe
c. Variables predictoras: (Constante), tiempo, pm, edad, pe
d. Variable dependiente: oxigen

Figura 2-160

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B		Estadísticos de colinealidad	
		B	Error típ.	Beta			Límite inferior	Límite superior	Tolerancia	FIV
1	(Constante)	102,934	12,403		8,299	,000	77,335	128,534		
	edad	-,227	,100	-,222	-2,273	,032	-,433	-,021	,661	1,513
	pd	-,022	,066	-,031	-,326	,747	-,158	,115	,706	1,416
	pe	-,370	,120	-,711	-3,084	,005	-,617	-,122	,119	8,437
	peso	-,074	,055	-,116	-1,359	,187	-,187	,038	,866	1,155
	pm	,303	,136	,522	2,221	,036	,022	,585	,114	8,744
	tiempo	-2,629	,385	-,685	-6,835	,000	-3,422	-1,835	,629	1,591
2	(Constante)	102,204	11,979		8,532	,000	77,532	126,876		
	edad	-,220	,096	-,215	-2,300	,030	-,416	-,023	,697	1,436
	pe	-,373	,117	-,719	-3,188	,004	-,615	-,132	,120	8,359
	peso	-,072	,053	-,113	-1,356	,187	-,182	,037	,875	1,143
	pm	,305	,134	,525	2,277	,032	,029	,581	,115	8,731
	tiempo	-2,683	,341	-,699	-7,867	,000	-3,385	-1,980	,771	1,297
3	(Constante)	98,148	11,786		8,328	,000	73,922	122,374		
	edad	-,198	,096	-,193	-2,068	,049	-,394	-,001	,717	1,395
	pe	-,348	,117	-,670	-2,963	,006	-,590	-,107	,123	8,147
	pm	,271	,134	,465	2,024	,053	-,004	,545	,119	8,418
	tiempo	-2,768	,341	-,721	-8,127	,000	-3,468	-2,068	,798	1,253

a. Variable dependiente: oxigen

Figura 2-161

En la Figura 2-159 se observan los tres pasos para llegar a la solución y las variables excluidas en cada paso. En la Figura 2-160 se observan los R^2altos, las tablas ANOVA de la regresión en cada paso (p-valores de la F muy pequeños que indican significatividad conjunta de los parámetros estimados muy buena) y el estadístico de Durbin-Watson en el tercer paso con valor 1,791 muy próximo al valor ideal 2 (no habrá problemas de autocorrelación. En la Figura 2-161 se observan los coeficientes estimados del modelo siendo todos significativos individualmente en el tercer paso (que contiene la regresión definitiva) con p-valores pequeños. La ecuación de ajuste será:

$$Oxigen = 98{,}148 - 0{,}198\ edad - 0{,}348\ pe + 0{,}271\ pm - 2{,}768\ tiempo$$

CAPÍTULO 3

AUTOCORRELACIÓN, HETEROSCEDASTICIDAD, MULTICOLINEALIDAD, NO LINEALIDAD Y NORMALIDAD

MODELOS CON AUTOCORRELACIÓN

En un modelo lineal de regresión múltiple, cuando la covarianza entre dos términos del error del modelo referidos a distintos momentos del tiempo es no nula, decimos que existe autocorrelación. El problema fundamental cuando existe autocorrelación en un modelo radica en que los estimadores MCO no son eficientes.

Partimos del modelo lineal:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_t \end{pmatrix} = \begin{pmatrix} 1\ X_{11}\ X_{21} \cdots X_{k1} \\ 1\ X_{12}\ X_{22} \cdots X_{k2} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ 1\ X_{1T}\ X_{2T} \cdots X_{kT} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_t \end{pmatrix}$$

que, abreviadamente, podemos representar mediante $Y = XB + u$, suponiendo una serie de hipótesis entre las que se encontraban que la variable u (término de error) es una variable aleatoria con esperanza nula ($E(u) = 0$) y matriz de covarianzas constante y diagonal ($Var(u) = \sigma^2 I_k$ matriz escalar). Es decir, que para todo t, la variable u_t tiene media cero y varianza σ^2 no dependiente de t, y además $Cov(u_i, u_j) = 0$ para todo i y para todo j distintos entre sí, pudiendo escribir $Var(u) = \sigma^2 I_k$.

El hecho de que $Cov(u_i, u_j) = 0$ para todo i distinto de j se denomina hipótesis de no autocorrelación. En este apartado estudiaremos el modelo lineal cuando esta hipótesis no se cumple, es decir, cuando existe *autocorrelación* o *correlación serial*.

Si se relaja la hipótesis $Var(u) = \sigma^2 I_k$ de modo que $Var(u) = V$, siendo V cualquier matriz, los parámetros estimados del modelo lineal resultan ser:

$$\hat{B} = (X'V^{-1}X)^{-1}X'V^{-1}Y \text{ con } E(\hat{B}) = B \text{ y } \Sigma(\hat{B}) = (X'V^{-1}X)^{-1}$$

No olvidemos que cuando se cumplían las hipótesis del modelo lineal:

$$\hat{B} = (X'X)^{-1}X'Y \text{ con } E(\hat{B}) = B \text{ y } \Sigma(\hat{B}) = \sigma^2 (X'X)^{-1}$$

Por tanto, en presencia de autocorrelación será necesario estimar los elementos de la matriz de varianzas covarianzas residual V. Esta tarea suele simplificarse suponiendo que las perturbaciones aleatorias del modelo siguen un determinado esquema de comportamiento que reduce el número de parámetros a estimar. Los esquemas más típicos son:

Modelo autorregresivo de orden 1 AR(1) $\rightarrow u_t = \rho u_{t-1} + e_t$
Modelo autorregresivo de orden 2 AR(2) $\rightarrow u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + e_t$
Modelo de medias móviles de orden 1 MA(1) $\rightarrow u_t = e_t + \rho e_{t-1}$

En general, las perturbaciones aleatorias pueden seguir modelos autorregresivos de medias móviles de cualquier orden, pero en el trabajo aplicado suele ser el modelo AR(1) el más utilizado, en cuyo caso tenemos:

$$u_t = \rho u_{t-1} + e_t \qquad \sigma_u^2 = \sigma_e^2/(1-\rho^2)$$

$$E(u) = 0,\ V = E(uu') = \sigma_u^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & & \rho^{T-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{bmatrix} = \sigma_u^2 \Omega$$

Con lo que ya conocemos V para poder estimar el modelo lineal mediante:

$$\hat{B} = (X'V^{-1}X)^{-1}X'V^{-1}Y \text{ con } E(\hat{B}) = B \text{ y } \Sigma(\hat{B}) = (X'V^{-1}X)^{-1}$$

DETECCIÓN DE LA AUTOCORRELACIÓN

Para analizar la autocorrelación de un modelo suele comenzarse por el análisis gráfico de los residuos, siendo esencial la gráfica de los residuos (a poder ser estudentizados) respecto del índice tiempo, que debe de presentar una estructura aleatoria libre de tendencia. También se pueden graficar los residuos u_t respecto de un retardo suyo u_{t-1} y si la mayoría de los puntos caen en el primer y tercer cuadrante hay indicios de una autocorrelación positiva, siendo la autocorrelación negativa cuando los puntos caen en el segundo y cuarto cuadrante.

Aparte del análisis gráfico es necesario realizar contrastes formales de autocorrelación, entre los que destacan Durbin Watson, Wallis, h-Durbin, Breusch-Godfrey y Cochrane-Orcutt.

Estadístico DW de Durbin-Watson

Consideramos el estadístico DW definido como sigue:

$$DW = \frac{\sum_{t=2}^{T}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T}\hat{u}_t^{\,2}} \cong 2(1-\rho) \Rightarrow \begin{cases} DW \cong 2 & si \quad \rho = 0 \\ DW \cong 0 & si \quad \rho = 1 \\ DW \cong 4 & si \quad \rho = -1 \end{cases}$$

Se puede adoptar la regla no demasiado rigurosa de que si DW vale 0 hay autocorrelación perfecta positiva; si DW se aproxima a 2 no hay autocorrelación y si DW se aproxima a 4 hay autocorrelación perfecta negativa. No obstante, DW se encuentra tabulado, y según la franja en la que caiga su valor, se acepta o rechaza la hipótesis de autocorrelación. En la tabla de DW elegimos la columna relativa a k (número de regresores en el modelo) y la fila relativa a T (tamaño muestral), lo que nos da los valores d_L y d_U. Se tiene:

- $DW < d_L \Rightarrow$ se rechaza $\rho = 0$ y se acepta $\rho > 0$
- $DW > 4 - d_L \Rightarrow$ se rechaza $\rho = 0$ y se acepta $\rho < 0$
- $d_U < DW < 4 - d_U \Rightarrow$ se acepta $\rho = 0$
- $4 - d_U < DW < 4 - d_L$ ó $d_L < DW < d_U \Rightarrow$ indeterminación

autocorrelación positiva	?	no autocorrelación	?	autocorrelación negativa
	d_L d_U		$4-d_U$ $4-d_L$	

El estadístico de Durbin Watson no debe utilizarse para modelos que introducen retardos en la variable dependiente ni para modelos sin término constante.

Estadístico h de Durbin

Supongamos ahora que el modelo presenta retardos de la variable dependiente:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \ldots \beta_r X_{rt} + \beta_i Y_{t-1} + \ldots + \beta_s Y_{t-s}$$

Consideramos el estadístico h de Durbin definido como sigue:

$$h=\hat{\rho}\sqrt{\frac{T}{1-T\hat{V}(\hat{\beta}_i)}}=\frac{\sum_{t=2}^{T}(\hat{u}_t\hat{u}_{t-1})}{\sum_{t=2}^{T}\hat{u}_{t-1}^{\,2}}\sqrt{\frac{T}{1-T\hat{V}(\hat{\beta}_i)}}\rightarrow N(0,1)$$

- Si $|h|<1{,}96$ se acepta la hipótesis $\rho=0$ frente a $\rho\neq 0$ (no existe autocorrelación)
- Si $|h|>1{,}96$ se rechaza la hipótesis $\rho=0$ al 95% (existe autocorrelación)

Estadístico D_4 de Wallis (datos trimestrales)

Se trata de una variante del estadístico DW de Durbin-Watson para el caso de series temporales de datos trimestrales. Consideramos el estadístico DW definido como sigue:

$$D_4=\frac{\sum_{t=2}^{T}(\hat{u}_t-\hat{u}_{t-4})^2}{SCE}=\frac{\sum_{t=5}^{T}(\hat{u}_t-\hat{u}_{t-4})^2}{\sum_{t=1}^{T}\hat{u}_t^{\,2}}$$

En la tabla de D_4 elegimos la columna relativa a k (número de regresores en el modelo) y la fila relativa a T (tamaño muestral), lo que nos da los valores d_{4L} y d_{4U}. Tenemos:

- $DW < d_{4L} \Rightarrow$ se rechaza $\rho = 0$ y se acepta $\rho > 0$
- $DW > 4- d_{4L} \Rightarrow$ se rechaza $\rho = 0$ y se acepta $\rho < 0$
- $D_{4U} < DW < 4- d_{4U} \Rightarrow$ se acepta $\rho = 0$
- $4-d_{4U} <DW<4- d_{4L}$ ó $d_{4L} < DW< d_{4U} \Rightarrow$ indeterminado

Razón de Von Neumann

Consideramos el estadístico ν definido como sigue:

$$\nu=\frac{\sum_{t=2}^{T}(\hat{u}_t-\hat{u}_{t-1})^2/T-1}{\sum_{t=1}^{T}\hat{u}_t^{\,2}/T}\Rightarrow t=\frac{\nu-E(\nu)}{\sigma(\nu)}\rightarrow N(0,1)\ \ con\ \begin{cases}E(\nu)=\dfrac{2-T}{T-1}\\ \sigma^2(\nu)=\dfrac{4T^2(T-2)}{(T+1)(T-1)^3}\end{cases}$$

- Si $|t|<1{,}96$ se acepta la hipótesis $\rho = 0$ frente a $\rho \neq 0$
- Si $|h|>1{,}96$ se rechaza la hipótesis $\rho = 0$ (al 95%)

Prueba de Breusch-Godfrey

Supongamos que el modelo puede presentar retardos de la variable dependiente, esquemas autorregresivos AR(p) de orden $p\geq 1$ en los residuos ($u_t = \rho_1 u_{t-1}+ \rho_2 u_{t-2} + \ldots+ \rho_p u_{t-p}+ \varepsilon_t$) o esquema de medias móviles MA(q) de orden q en los residuos ($u_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}+ \theta_2 \varepsilon_{t-2} + \ldots+ \theta_q \varepsilon_{t-q}$), o un esquema ARMA($p,q$) en los residuos ($u_t = \rho_1 u_{t-1}+ \rho_2 u_{t-2} + \ldots+ \rho_p u_{t-p}+ \varepsilon_t+ \theta_1 \varepsilon_{t-1}+ \theta_2 \varepsilon_{t-2} + \ldots+ \theta_q \varepsilon_{t-q}$). El modelo puede tener la forma:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \ldots\ \beta_r X_{rt} + \beta_i Y_{t-1} + \ldots + \beta_s Y_{t-s}+ u_t$$

El contraste de Breusch-Godfrey para contrastar la autocorrelación se realiza mediante los siguientes pasos:

1º) Se estima por MCO el modelo original y se obtienen los residuos estimados $\hat{u}_t$.

2º) Se estima por MCO la regresión de $\hat{u}_t$ sobre los regresores originales del modelo y p retardos de los residuos $\hat{u}_{t-1},\hat{u}_{t-2},\cdots,\hat{u}_{t-p}$, en caso de un esquema AR(p) en los residuos. Se obtiene el R^2 de esta regresión.

3º) Para tamaño muestral T suficientemente grande se cumple que $(T-p)R^2 \rightarrow \chi^2_p$. Éste será el estadístico del contraste de Breusch-Godfrey. Para valores muestrales de este estadístico mayores que el valor crítico de la Chi-cuadrado con p grados de libertad se acepta la presencia de autocorrelación con retardo de orden p en los residuos (esquema residual AR(p)).

El problema en este método es la detección del orden autorregresivo p de los residuos. Para identificar adecuadamente p se usan de forma conjunta las funciones de autocorrelación FAC y autocorrelación parcial FACP de los residuos. Para residuos AR(p) la FAC decrece exponencialmente, es sinusoidal o con alternancia de signos, mientras que en la FACP sólo existen p valores significativamente distintos de cero. Para un esquema residual MA(q) es la FAC la que tiene sólo los q primeros valores significativamente distintos de cero, mientras que la FACP presenta decrecimiento exponencial, alternancia de signos o es sinusoidal. En un esquema residual ARMA(p,q) se combinan un AR(p) y un MA(q).

SOLUCIONES PARA LA AUTOCORRELACIÓN

La presencia de autocorrelación en un modelo suele solventarse mediante el método de Cochrane-Orcutt o mediante la introducción de variables *dummy* adecuadas en el modelo. Existen otros métodos menos utilizados como el método de estimación de Durbin y el procedimiento de Prais-Winsten.

Método de Mínimos Cuadrados Generalizados

Este método se basa en realizar la estimación del modelo con autocorrelación mediante mínimos cuadrados generalizados MCG con la matriz V siguiente:

$$V=E(uu')=\sigma_u^2\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & & \rho^{T-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{bmatrix}=\sigma_u^2\Omega \Rightarrow$$

$$\Omega^{-1}=\frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & 0 & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

No olvidemos que $u_t=\rho u_{t-1}+e_t$, $\sigma_u^2=\sigma_e^2/(1-\rho^2)$ y $\hat{B}=(X'V^{-1}X)^{-1}X'V^{-1}Y$, siendo $E(\hat{B})=B$ y $\Sigma(\hat{B})=(X'V^{-1}X)^{-1}$. También se cumple $\hat{B}=(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$.

Método de Prais-Winsten

Alternativamente, se obtiene el estimador MCG estimando por MCO el modelo transformado: $Y_t-\rho Y_{t-1}=\beta_1(1-\rho)+\beta_2(X_{2t}-\rho X_{2t-1})+\ldots+\beta_k(X_{kt}-\rho X_{kt-1})$ para $t=2,3,\ldots,T$. Para $t=1$ se hace la transformación:

$$\sqrt{1-\rho^2}Y_1=\sqrt{1-\rho^2}(\beta_1+\beta_2X_{21}+\cdots+\beta_kX_{k1}+u_1).$$

Método iterativo de Cochrane-Orcutt

Consideremos el modelo $Y_t=\beta_0+\beta_1X_{1t}+\ldots+\beta_kX_{kt}+u_t$ con esquema AR(1) en los residuos: $u_t=\rho u_{t-1}+e_t$.

En el primer paso del método iterativo estimamos el modelo inicial mediante MCO y hallamos los residuos estimados $\hat{u}_t$. A continuación estimamos ρ por MCO mediante la regresión $\hat{u}_t=\hat{\rho}\hat{u}_{t-1}+\upsilon_t$ y se plantea el modelo:

$$\underbrace{Y_t-\hat{\rho}Y_{t-1}}_{Y_t^*}=\underbrace{\beta_1(1-\hat{\rho})}_{\beta_1^*}+\beta_2^*\underbrace{(X_{2t}-\hat{\rho}X_{2t-1})}_{X_{2t}^*}+\ldots+\beta_k^*\underbrace{(X_{kt}-\hat{\rho}X_{kt-1})}_{X_{kt}^*}+u_t^*$$

Estimamos el modelo anterior por MCO obteniendo:

$$Y_t^*=\hat{\beta}_1^*+\hat{\beta}_2^*X_{2t}^*+\ldots+\hat{\beta}_k^*X_{kt}^*+\hat{u}_t^*$$

En el segundo paso del método iterativo estimamos ρ^* por MCO mediante la regresión $\hat{u}_t^*=\hat{\rho}^*\hat{u}_{t-1}^*+w_t$ y realizamos las misma operativa que en el primer paso planteando el modelo:

$$\underbrace{Y_t^*-\hat{\rho}Y_{t-1}^*}_{Y_t^{**}}=\underbrace{\beta_1^*(1-\hat{\rho}^*)}_{\beta_1^*}+\beta_2^{**}\underbrace{(X_{2t}^*-\hat{\rho}^*X_{2t-1}^*)}_{X_{2t}^{**}}+\ldots+\beta_k^{**}\underbrace{(X_{kt}^*-\hat{\rho}^*X_{kt-1}^*)}_{X_{kt}^{**}}+u_t^{**}$$

Estimamos el modelo anterior por MCO obteniendo:

$$Y_t^{**}=\hat{\beta}_1^{**}+\hat{\beta}_2^{**}X_{2t}^{**}+\ldots+\hat{\beta}_k^{**}X_{kt}^{**}+\hat{u}_t^{**}$$

En el tercer paso del método iterativo estimamos ρ^{**} por MCO mediante la regresión $\hat{u}_t^{**}=\hat{\rho}^{*}*\hat{u}_{t-1}^{**}+\lambda_t$ y realizamos las misma operativa que en el primer paso anterior. Las estimaciones finales de los parámetros del modelo serán las obtenidas en el último paso.

El proceso finalizará cuando la diferencia entre dos estimaciones consecutivas de ρ sea tan pequeña como queramos o cuando superemos un determinado número de iteraciones.

Método de Durbin

Este método se basa en estimar ϕ (parámetro correspondiente a Y_{t-1}) en la ecuación:

$$Y_t=b_1(1-\phi)+b_2X_{2,t}-b_2\phi X_{2,t-1}+\ldots+b_kX_{k,t}-b_k\phi X_{k,t-1}+\phi Y_{t-1}+u_t \quad t=2,3,\ldots T$$

Una vez estimado el valor de ϕ, se aplica el método de Cochrane-Orcutt para estimar el modelo con autocorrelación tomando $\rho=\phi$.

MODELOS CON HETEROSCEDASTICIDAD

Entre las hipótesis a cumplir en el modelo lineal con series temporales suponíamos que la variable ε (término de error) es una variable aleatoria con esperanza nula y matriz de covarianzas constante y diagonal (matriz escalar). Es decir que, para todo instante temporal t, la variable ε_t tiene media cero y varianza σ^2 no dependiente de t, y además $Cov(\varepsilon_i, \varepsilon_j) = 0$ para todo i y para todo j instantes temporales distintos entre sí. El hecho de que la varianza de ε_t sea constante para todo t (que no dependa de t), se denomina hipótesis de *homoscedasticidad condicional* y puede también expresarse como $V(\varepsilon|X_1,X_2, ..., X_k) = \sigma^2$ y $V(Y|X_1,X_2, ..., X_k) = \sigma^2$.

La importancia del incumplimiento de la hipótesis de homoscedasticidad condicional radica, entre otras cosas, en que los estimadores obtenidos por MCO no son de varianza mínima aunque sigan siendo insesgados. Además, para cada variable del modelo se estimará una varianza del error.

DETECCIÓN DE LA HETEROSCEDASTICIDAD

Para analizar la heteroscedasticidad condicional de un modelo suele comenzarse por el análisis gráfico de los residuos, siendo esenciales las gráficas de los residuos (a poder ser estudentizados) respecto de la variable endógena predicha y respecto de las exógenas, que deben de presentar una estructura aleatoria libre de tendencia. El gráfico de los residuos contra cada variable exógena permite detectar como *variable más culpable de heteroscedasticidad* aquella cuyo gráfico se separa más de la aleatoriedad. También es un instrumento gráfico útil la gráfica de valores observados contra valores predichos, cuyos puntos han de ser lo más ajustados posible a la diagonal del primer cuadrante.

Aparte del análisis gráfico es necesario realizar contrastes formales de heteroscedasticidad, entre los que destacan Goldfeld-Quandt, Glesjer, Breush-Pagan, White, y RESET de Ramsey.

Contraste de Goldfeld-Quandt

Se detecta la variable culpable de heteroscedasticidad X_j y se ordenan las observaciones muestrales de todas las variables en función de la ordenación de X_j. Se suprimen las c observaciones centrales de modo que sean 1/3 del total y se hacen dos regresiones distintas con lo que queda. Se considera el estadístico $F = SCE1/SCE2$ que, bajo la hipótesis de homoscedasticidad, es una F de Fisher con $(T - c - 2k)/2$ grados de libertad en ambas dimensiones. $SCE1$ es la suma de cuadrados del error de la regresión relativa a la submuestra de valores bajos de X_j, y $SCE2$ es la relativa a la submuestra de los valores altos.

Contraste de Glesjer

Se estiman los residuos del modelo u_t por MCO y se realiza la regresión:

$$|u_t| = d_0 + d_1 Z_i^h + r_i \text{ , } Z_i = \text{variable culpable, } h = 1, -1 \text{ ó } 1/2$$

Contrastar d_1=0 en el modelo anterior es equivalente a contrastar la hipótesis de homoscedasticidad en el modelo inicial.

Contraste W de White

Se efectúa una regresión de los cuadrados de los residuos MCO sobre todas las variables independientes del modelo, sus cuadrados y sus productos cruzados de dos en dos. La homoscedasticidad se acepta si la F permite aceptar la nulidad conjunta de todos los parámetros salvo la constante.

Contraste RESET de Ramsey

En una primera etapa se estiman los residuos u_t del modelo inicial y los correspondientes valores ajustados de Y_t. Para cada t se calculan las m primeras potencias de las estimaciones de Y_t.

En una segunda etapa se realiza la regresión auxiliar de las estimaciones de u_t contra todos los regresores del modelo inicial y las m primeras potencias de las estimaciones de Y_t. Si no hay significatividad de los coeficientes hay heteroscedasticidad.

Contraste de Breush-Pagan

Se utiliza cuando la varianza no constante de las perturbaciones puede expresarse como $\sigma_i^2 = h(Z_i'a)$ siendo h una función y Z_i un vector con las variables que producen la heteroscedasticidad. Si $a = 0$ la varianza es constante.

El contraste se realiza obteniendo los residuos del ajuste del modelo y considerando los cocientes g_i entre los cuadrados de cada residuo del modelo original y la media aritmética de dichos cuadrados (o la estimación MV de la varianza del error). Al realizar la regresión entre las variables g_i y las variables culpables Z_i (p-1 variables) más la constante, la cantidad $Q = VE/2$ se distribuye según una Chi-cuadrado con p-1 grados de libertad bajo la hipótesis nula de homoscedasticidad ($a = 0$). VE es la varianza explicada de la regresión anterior.

La mejor propiedad de este contraste es que no es necesario explicitar la forma de la función h.

Heteroscedasticidad condicional. Contrastes ARCH y GARCH

Dado que la varianza del término del error puede interpretarse como una incertidumbre en la relación económica especificada en un modelo, una mayor incertidumbre en el pasado podría afectar al comportamiento presente, por lo que Engel (1982) propuso que la matriz de varianzas y covarianzas del término del error del modelo $Y = X\,B + u$ depende del cuadrado de los errores observados en el pasado: $\sigma_t^2 = V(u_t) = d_0+d_1\, u_{t-1}{}^2+\ldots+ d_p\, u_{t-p}{}^2$. Evidentemente, si $d_1 =\ldots= d_p = 0$ no existe efecto ARCH (no hay heteroscedasticidad)

En la práctica, para realizar el contraste ARCH(p), se comienza estimando los residuos u_t del modelo inicial y después se realiza la regresión auxiliar $u_t^2 = d_0+d_1\, u^2_{t-1}+\ldots+ d_p u^2_{t-p}$. La homoscedasticidad existe si la F acepta la nulidad conjunta de los parámetros salvo la constante.

Un problema habitual en este método es identificar p, para lo cual suele usarse el correlograma estimado de los residuos al cuadrado, tomando p como el retardo a partir del cual los coeficientes son no significativos

Contrastes GARCH

Boliersiev (1986) amplió la estructura ARCH de Engel incluyendo en la especificación de la varianza valores retardados de la misma, de modo que en una estructura GARCH(p,q) tenemos: $\sigma_t^2 = V(u_t) = d_0+d_1u^2_{t-1}+\ldots+ d_pu^2_{t-p}+\alpha_1\sigma^2_{t-1} +\ldots+ \alpha_q\sigma^2_{t-q}$. La identificación de p y q se realiza como en los modelos ARIMA a través del correlograma y el correlograma estimado. Actualmente se utilizan otras estructuras GARCH más amplias, como los modelos TARCH, EGARCH, PARCH, CGARCH y GARCH-M.

Contraste de rangos

Se trata de un método no paramétrico que se resume en los siguientes pasos:

- Se aplican mínimos cuadrados al modelo original.
- Se calculan los rangos de los residuos en valor absoluto y de las observaciones de la variable culpable de la heteroscedasticidad.
- Se calcula el coeficiente de correlación por rangos de Sperman $r_s = 1-6\sum D_i^2/(T(T^2-1))$ siendo D_i la diferencia entre los rangos de la variable explicativa y $|u_i|$.

- Se contrasta la hipótesis nula H_0: $\rho_s = 0$ mediante el estadístico $r_s(T-2)^{1/2}/(1-r_s^2)^{1/2}$ que sigue una t de Student con T-2 grados de libertad. Si se acepta H_0 hay homoscedasticidad, y si se rechaza, hay heteroscedasticidad.

SOLUCIONES PARA LA HETEROSCEDASTICIDAD

Mínimos Cuadrados Generalizados y Mínimos Cuadrados Ponderados

En general, para resolver el problema de heteroscedasticidad es necesario realizar la estimación por MCG (Mínimos Cuadrados Generalizados). Pero si la estructura de la varianza de las perturbaciones es conocida, se facilita el cálculo de los estimadores. Si se puede suponer aproximadamente que $\sigma_i^2 = f(Z_i)$, siendo Z_i un vector de variables que incluye una o varias variables exógenas de la regresión y f una función cualquiera, entonces puede reducirse la estimación MCG a MCO (Mímimos Cuadrados Ordinarios) con ponderaciones adecuadas (Mínimos Cuadrados Ponderados).

Entre las estructuras típicas de la varianza tenemos $\sigma_i^2 = \sigma^2X_{ji}$, $\sigma_i^2 = \sigma^2X_{ji}^2$, $\sigma_i^2 = a+bX_{ji}$ y $\sigma_i^2 = exp(Z_i'a)$, siendo las dos primeras las más comunes y la tercera una translación de la primera. En los dos primeros casos las matrices de varianzas covarianzas resultan ser:

$$\Sigma = \begin{bmatrix} X_{j1} & 0 & \cdots & 0 \\ 0 & X_{j2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & X_{jT} \end{bmatrix} \quad y \quad \Sigma = \begin{bmatrix} X_{j1}^2 & 0 & \cdots & 0 \\ 0 & X_{j2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & X_{jT}^2 \end{bmatrix}$$

En estos casos la regresión MCG coincide con la MCO utilizando como ponderaciones los valores $1/X_{ji}$ y $1/X_{ji}^2$, o sea los inversos de los elementos de la diagonal de Σ.

Si la estructura de la varianza es $\sigma_i^2 = \sigma^2X_{ji}$, el modelo se transforma dividiendo sus términos por $X_{ji}^{1/2}$, de modo que estimaríamos por MCO el modelo:

$$\frac{Y_t}{\sqrt{X_{jt}}} = \frac{\beta_1}{\sqrt{X_{jt}}} + \frac{\beta_2}{\sqrt{X_{jt}}} + \cdots + \frac{\beta_k}{\sqrt{X_{jt}}} + \frac{u_t}{\sqrt{X_{jt}}}$$

En general, para resolver el problema de heteroscedasticidad es conveniente tomar logaritmos. También pueden suprimirse las variables más culpables con justificación estadística y económica o introducir variables *dummy* adecuadas.

Para detectar la mejor forma funcional que sigue la varianza, se ajustan distintos modelos para las distintas formas funcionales del tipo siguiente:

$$|\hat{u}_t| = a + b\sqrt{X_{jt}} + e_t$$
$$|\hat{u}_t| = a + bLnX_{jt} + e_t$$
$$|\hat{u}_t| = a + b\frac{1}{X_{jt}} + e_t$$

Se toma como esquema de heteroscedasticidad aquella forma funcional para la que la regresión correspondiente presenta un R^2 más elevado.

Ajuste de White

El método de White permite estimar un modelo en presencia de heteroscedasticidad desconocida, es decir, sin necesidad de especificar ningún tipo de comportamiento conocido para la varianza de las perturbaciones. Este método proporciona estimaciones consistentes para los parámetros y para la matriz de varianzas covarianzas.

Soluciones para la heteroscedasticidad condicional: modelos ARCH y GARCH

Detectada una estructura de heteroscedasticidad autorregresiva condicional **ARCH(*p*)** en un modelo, existen métodos para su estimación. La estructura del modelo a estimar será:

$$Y = XB + u_t$$
$$\sigma_t^2 = V(u_t) = d_0 + d_1\, u_{t-1}{}^2 + \ldots + d_p\, u_{t-p}{}^2$$

La estructura anterior puede ampliarse a la estimación de un modelo **GARCH(*p,q*)** con heteroscedasticidad autorregresiva condicional generalizada. El modelo a estimar en este caso es:

$$Y = XB + u_t$$
$$\sigma_t^2 = d_0 + d_1 u^2_{t-1} + \ldots + d_p u^2_{t-p} + \alpha_1 \sigma^2_{t-1} + \ldots + \alpha_q \sigma^2_{t-q}.$$

La estructura **GARCH-M** amplía la estructura GARCH introduciendo la varianza como término de la ecuación el modelo. El modelo a estimar en este caso será:

$$Y = XB + \lambda\sigma_t^2 + u_t$$
$$\sigma_t^2 = d_0 + d_1 u^2_{t-1} + \ldots + d_p u^2_{t-p} + \alpha_1 \sigma^2_{t-1} + \ldots + \alpha_q \sigma^2_{t-q}.$$

Dos variantes de la estructura GARCH-M consisten en introducir en la ecuación del modelo el logaritmo de la varianza (en lugar de la varianza) o incluso la desviación típica:

$$Y = XB + \lambda log(\sigma_t^2) + u_t$$
$$Y = XB + \lambda\sigma_t + u_t$$

La estructura GARCH(*p,q*) también puede ampliarse con la introducción de regresores exógenos en el estructura de la varianza. El modelo a estimar en este caso es:

$$Y = XB + u_t$$
$$\sigma_t^2 = d_0 + d_1 u^2_{t-1} + \ldots + d_p u^2_{t-p} + \alpha_1 \sigma^2_{t-1} + \ldots + \alpha_q \sigma^2_{t-q}. + Z\pi$$

La estructura **TARCH(*p,q,r*)** denominada *Threshold* GARCH también es una ampliación del modelo GARCH(*p,q*). El modelo a estimar en este caso es:

$$Y = XB + u_t$$
$$\sigma_t^2 = d_0 + d_1 u^2_{t-1} + \ldots + d_p u^2_{t-p} + \alpha_1 \sigma^2_{t-1} + \ldots + \alpha_q \sigma^2_{t-q} + \beta_1 u^2_{t-1}\Gamma_{t-1} + \ldots + \beta_r u^2_{t-r}\Gamma_{t-r}$$
($\Gamma_t = 1$ si $u_t < 0$ y 0 en otro caso)

La estructura **EGARCH(*p,q,r*)** denominada *Exponential* GARCH también es una ampliación del modelo GARCH(*p,q*). El modelo a estimar en este caso es:

$$Y = XB + u_t$$
$$\sigma_t^2 = d_0 + d_1 |u_{t-1}/\sigma_{t-1}| + \ldots + d_p |u_{t-p}/\sigma_{t-p}| + \alpha_1 log(\sigma^2_{t-1}) + \ldots + \alpha_q\, log(\sigma^2_{t-q}) + \beta_1\, u_{t-1}/\sigma_{t-1} + \ldots + \beta_r\, u_{t-r}/\sigma_{t-r}$$

La estructura **PARCH(*p,q,r*)** denominada *Power* ARCH también es una ampliación del modelo GARCH(*p,q*). El modelo a estimar en este caso es:

$$Y = XB + u_t$$
$$\sigma_t^r = d_0 + d_1(|u_{t-1}| - \beta_1 u_{t-1})^r + \ldots + d_p(|u_{t-p}| - \beta_p u_{t-p})^r + \alpha_1 \sigma^r_{t-1} + \ldots + \alpha_q \sigma^r_{t-q}.$$
($r>0$, $|\beta_i| \leq 1$ $i=1,\ldots,s$, $\beta_i=0$ para $i>s$ y $s \leq p$)

La estructura **CGARCH(*p,q,r*)** denominada *Component* GARCH o *Component* ARCH también es una ampliación del modelo GARCH(*p,q*). El modelo a estimar en este caso es:

$$Y = XB + u_t$$
$$\sigma_t^2 = d + d_1(u^2_{t-1} - d) + d_p(u^2_{t-p} - d) + \alpha_1(\sigma^2_{t-1} - d) + \ldots + \alpha_q(\sigma^2_{t-q} - d)$$

MULTICOLINEALIDAD

En el modelo lineal $Y = XB + u$, suponíamos una serie de hipótesis entre las que se encontraban que las variables $X_1, X_2,..., X_k$ son linealmente independientes, es decir, no existe relación lineal exacta entre ellas. Esta hipótesis se denomina hipótesis de independencia, y cuando no se cumple, decimos que el modelo presenta *multicolinealidad.*

En caso de multicolinealidad, al tener fuerte asociación lineal entre las variables explicativas $X'X$ tendría determinante cercano a 0 y no sería calculable $(X'X)^{-1}$ con lo que no se podría hallar el vector de estimaciones de los parámetros $(X'X)^{-1} X'Y$.

Detección de la multicolinealidad

Como síntomas más comunes de multicolinealidad tenemos los siguientes:

- Valores altos en módulo en la matriz de correlaciones de las variables explicativas.
- Poca significatividad de las variables X y a la vez R^2 alto.
- Gran significatividad conjunta del modelo (gran rechazo de $R^2 = 0$).
- Influencia en las estimaciones de la eliminación de una observación en el conjunto de datos.
- ***Factores de inflación*** de la varianza $VIF = 1/(1-R_j^2)$ elevados (>10), donde R_j^2 es el R^2 de la regresión auxiliar de la variable explicativa j en función de las demás variables explicativas.
- Valores propios λ_i de $X'X$ cercanos a cero o ***índice de condición*** $(\lambda_{máx} / \lambda_{mín})^{1/2}$ mayor que 30.
- El ***contraste de Farrar-Glauber*** basado en el estadístico $G = -[T-1-(2k+5)/6]Ln|R_{xx}|$ que bajo la hipótesis nula de no multicolinealidad es una Chi-cuadrado con $k(k-1)/2$ grados de libertad. T es el tamaño muestral, k-1 el número de variables explicativas y R_{xx} su matriz de correlaciones.

Soluciones para la multicolinealidad

Entre las soluciones más comunes para la multicolinealidad tenemos:

- Ampliar la muestra o transformar las variables (por ejemplo a ratios o diferencias).
- Suprimir algunas variables con justificación estadística y económica.

- Sustitución de las variables explicativas por sus componentes principales más significativas (puntuaciones).
- Utilizar el modelo en diferencias vigilando la autocorrelación.
- Usar la *regresión en cadena*, que ofrece como estimadores de los parámetros $(X'X+cI)^{-1} X'Y$ siendo c una constante adecuada. La matriz de varianzas covarianzas adopta la forma $\sigma^2(X'X+cI)^2X'X(X'X+cI)^{-1}$. En la práctica suele tomarse c como un valor entre 0,01 y 0,1 que hace que el ajuste sea bueno en cuanto a R^2 y significatividad individual y conjunta.

NORMALIDAD RESIDUAL

El problema de la falta de normalidad en los residuos

Una de las hipótesis importantes a cumplir en el modelo de regresión múltiple es la normalidad de los residuos. Aunque dicha hipótesis no es necesaria para la obtención de los estimadores de los parámetros del modelo de regresión por el método de los mínimos cuadrados ordinarios, sí es estrictamente necesaria para la realización de la inferencia en el modelo.

Para probar la normalidad de los residuos puede utilizarse cualquier contraste de ajuste a una distribución normal, por ejemplo el contraste de la Chi-cuadrado o el contraste de Kolmogorov-Smirnov. No obstante, existen también contrastes específicos para comprobar el ajuste de un conjunto de datos a una distribución normal, como por ejemplo contraste de normalidad de Shapiro y Wilks y los contrastes de normalidad de asimetría, curtosis y Jarque-Bera.

Contraste de normalidad de Shapiro y Wilks

El contraste de Shapiro y Wilks mide el ajuste de los residuos de la regresión a una recta al dibujarla en un papel probabilístico normal. Se rechaza la normalidad cuando el ajuste es bajo, que corresponde a valores pequeños del estadístico del test. Dicho estadístico toma la expresión:

$$w = \frac{1}{ns^2}\left[\sum_{j=1}^{h} a_{j,n}(x_{(n-j+1)} - x_{(j)})\right]^2 = \frac{A^2}{ns^2}$$

donde $ns^2=\sum(x_i - \bar{x})^2$, h es $n/2$ si n es par y $(n-1)/2$ si n es impar. Los coeficientes $a_{j,n}$ están tabulados y $x_{(j)}$ es el valor ordenado en la muestra que ocupa el lugar j. La distribución de w está tabulada, y se rechaza la normalidad cuando su valor calculado a partir de la muestra es menor que el correspondiente valor crítico dado en las tablas. De todas formas, puede utilizarse el criterio del p-valor, rechazando la hipótesis nula de normalidad de los datos al nivel α cuando el p-valor es menor que α, y aceptándola en caso contrario.

Contrastes de normalidad de asimetría, curtosis y Jarque-Bera

Estos contrastes se basan en los coeficientes de asimetría y curtosis muestrales (la muestra son los residuos del modelo). Si la hipótesis de normalidad es cierta, el estadístico del contraste, que es el coeficiente de asimetría muestral $\alpha_1 = m_3 / m_2^{3/2}$, tiene una distribución asintóticamente normal de media cero y varianza $6/n$, siendo m_2 y m_3 los momentos muestrales centrados en la media de órdenes 2 y 3 respectivamente. Tenemos:

$$\alpha_1 = \frac{m_3}{m_2^{3/2}} \to N\left(0, \sqrt{\frac{6}{n}}\right)$$

Este estadístico α_1 permite contrastar la hipótesis de que los residuos provienen de una distribución con simetría normal (asimetría = 0) y se basa en que si la hipótesis de normalidad es cierta, el coeficiente de asimetría estima un parámetro de la población que es cero (el coeficiente de asimetría de una distribución normal es cero). Para realizar el contraste se halla el valor k tal que $P(\alpha_1 \geq k) = \alpha$, siendo α el nivel de significación establecido para el contraste. Si el valor del estadístico α_1 para los residuos es mayor que k se rechaza la hipótesis nula de simetría, y por supuesto la de normalidad.

De la misma forma, si la hipótesis de normalidad es cierta, el estadístico del contraste, que es el coeficiente de curtosis muestral $\alpha_2 = m_4 / m_2^2 - 3$, tiene una distribución asintóticamente normal de media cero y varianza $24/n$, siendo m_2 y m_4 los momentos muestrales centrados en la media de órdenes 2 y 4 respectivamente.

$$\alpha_2 = \frac{m_4}{m_2^2} - 3 \to N\left(0, \sqrt{\frac{24}{n}}\right)$$

Este estadístico α_2 permite contrastar la hipótesis de que los residuos provienen de una distribución con curtosis normal (curtosis = 0) y se basa en que si la hipótesis de normalidad es cierta, el coeficiente de curtosis estima un parámetro de la población que es cero (el coeficiente de curtosis de una distribución normal es cero). Para realizar el contraste se halla el valor k tal que $P(\alpha_2 \geq k) = \alpha$, siendo α el nivel de significación establecido para el contraste. Si el valor del estadístico α_2 para los residuos es mayor que k se rechaza la hipótesis nula de curtosis cero, y por supuesto la de normalidad.

Para muestras grandes, el contraste de Jarque-Bera usa los dos estadísticos anteriores mediante la consideración del estadístico de Bowman-Shelton siguiente:

$$B = n\left[\frac{\alpha_1^2}{6} + \frac{\alpha_2^2}{24}\right] \to \chi_2^2$$

Es posible utilizar para estos contrastes (como siempre) el criterio del p-valor, rechazando la hipótesis nula de normalidad de los residuos al nivel α cuando el p-valor es menor que α en alguno de ellos, y aceptándola cuando el p-valor es mayor que α en los dos.

Como criterio más suave sobre la normalidad, suele considerarse normal la población cuya muestra presenta coeficientes de asimetría y curtosis comprendidos entre –2 y 2.

Soluciones para la falta de normalidad en los residuos

Habitualmente, la falta de normalidad en los residuos suele provenir de la presencia de datos atípicos que generan una distribución más apuntada o no simétrica. Estos problemas en los residuos suelen aparecer cuando se omiten variables relevantes en el modelo o cuando existe falta de linealidad en la especificación del mismo. Si se arreglan previamente los problemas citados, suelen solucionarse los problemas de normalidad residual. Cuando los residuos no son normales por la presencia de más de una moda, los datos suelen provenir de varias poblaciones, lo que puede arreglarse con la introducción de variables ficticias en el modelo para las diferentes poblaciones. En otras ocasiones, la solución para la falta de normalidad es la transformación adecuada de las variables para conseguirla, por ejemplo la transformación de Box Cox y sus derivados.

NO LINEALIDAD Y ERRORES DE ESPECIFICACIÓN

La técnica de los mínimos cuadrados ordinarios MCO es el caballo de batalla de los económetras y se utiliza de modo rutinario en el análisis de una gran variedad de conjuntos de datos. Bajo los supuestos exigidos al modelo lineal, los estimadores mínimo cuadráticos poseen las propiedades deseables y, por ello, pueden emplearse con fiabilidad. Sin embargo, nos enfrentamos a una pregunta crucial. ¿Cómo saber si los supuestos que ocultan los MCO son válidos para un conjunto determinado de datos? ¿Cómo conocer las propiedades del término de perturbación no observable? ¿Cómo saber qué variables incluir en la matriz X y en qué forma funcional hacerlo? Cuando alguno de los supuestos subyacentes carece de validez, ¿qué sucede con los estimadores MCO? ¿Siguen siendo útiles o resultan confusos? ¿Existen estimadores y procedimientos de inferencia alternativos que resulten más apropiados bajo supuestos alternativos? En este capítulo responderemos a estas preguntas.

El error de especificación aparece cuando alguno de los supuestos está equivocado.

Ciertos errores de especificación tienen implicaciones menores; otros, sin embargo, las tienen muy graves. Resulta tremendamente importante estar alerta de posibles errores de especificación y verificar su presencia. En este capítulo se estudia cómo muchas veces es necesario utilizar y desarrollar especificaciones y procedimientos de inferencia más complejos que los que subyacen en la técnica de los MCO.

La especificación del modelo lineal se centra en el vector de términos de perturbaciones u y en la matriz X. Recordemos que los *supuestos esenciales del modelo lineal* $y = X\beta + u$ *relativos a las perturbaciones* son: u_i son indepedientes idénticamente distribuidas $N(0,\sigma^2)$ $i = 1, \ldots, n$ (perturbaciones normales de ruido blanco), homoscedasticidad ($E(uu') = diag[\sigma_u^2 \ldots \sigma_u^2]$ o matriz de varianzas covarianzas residual constante), endogeneidad ($E(X_{it}u_s) = 0$ para todo $i = 1,\ldots, k$ y $t, s = 1,\ldots, n$ o incorrelación entre las perturbaciones y las variables independientes) y ausencia de autocorrelación ($E(u_t u_{t-s}) = 0$ $(s \neq 0)$ o ausencia de correlaciones entre perturbaciones adyacentes). La heteroscedasticidad (ausencia de homoscedasticidad) es muy frecuente en aplicaciones con datos de corte transversal, aunque puede encontrarse también en aplicaciones con datos de series temporales. En las aplicaciones de series temporales se dan fuertes correlaciones entre perturbaciones adyacentes y, tal vez, correlaciones menores entre perturbaciones más alejadas entre sí (*autocorrelación*). De modo similar, y cuando trabajamos con datos de corte transversal, es posible que ciertas unidades compartan perturbaciones comunes.

También existen en el modelo lineal *supuestos relativos a la matriz X*, entre los que destacan su rango pleno (ausencia de multicolinealidad), la inclusión y no exclusión de variables relevantes en X (ausencia de errores de especificación en la selección de las variables explicativas) y problemas de especificación de una forma funcional incorrecta para el modelo.

Otro supuesto común es la *estacionariedad de las variables del modelo*. La mayoría de los procedimientos de inferencia tradicionales suponen que las variables son estacionarias. Cuando no se da este caso nos enfrentamos a procedimientos de inferencia no estándar y nos introducimos en el campo de las variables integradas, la cointegración, los modelos de corrección del error, etc., que se tratarán más adelante.

También pueden existir problemas de especificación con β. Las especificaciones del modelo lineal asumen de forma implícita que β es un vector constante, tanto en el conjunto de observaciones actuales como en otras observaciones muestrales posibles. Estamos ante la especificación de ausencia de cambio estructural que se tratará en un capítulo posterior.

Las prácticas econométricas habituales formulan un modelo basado en teoría o en anteriores descubrimientos econométricos, estiman los parámetros del modelo mediante los datos muestrales relevantes disponibles, y examinan los estimadores resultantes y estadísticos asociados con el fin de juzgar la validez del modelo especificado.

Dicho examen suele centrarse en el ajuste global, en la concordancia con los signos de unos coeficientes previamente supuestos, en la significación estadística de los coeficientes y en la comprobación de la autocorrelación de las perturbaciones. Si el modelo cumple dichos criterios satisfactoriamente, la nueva ecuación pasaba a engrosar la literatura de la materia y podría utilizarse para realizar predicciones con datos externos a la escala temporal o al rango empírico de la muestra. En caso de que el modelo se clasifique de insatisfactorio, el investigador seguirá intentando hallar la reformulación que cumpla los requisitos necesarios.

Actualmente existen tendencias basadas en comprobar de todas las formas posibles las especificaciones y que sólo deberán utilizarse aquellas que sobrevivan a este proceso de prueba y que correspondan a un modelo económico razonable. Así se llega a una auténtica batería de pruebas de diagnóstico que no pueden utilizarse ni de forma automática ni rutinaria, ya que requieren una dosis de juicio, intuición económica o sentido común. Algunos de los contrastes resaltan un error o errores de especificación en particular. Otros indican que determinada especificación no funciona bien sin señalar explícitamente un problema preciso. Finalmente, puede ocurrir que sobrevivan a este proceso de prueba o que algunas especificaciones superen un cierto tipo de pruebas estadísticas pero no otras.

Error de especificación en la selección de las variables explicativas

La especificación más importante del modelo lineal relativa a la matriz X es que sea una matriz no estocástica de rango pleno igual a k (ausencia de multicolinealidad). Pero puede haber posibles problemas adicionales con X, entre los que destacan:

1. ***Exclusión de variables relevantes (variables omitidas).*** La teoría económica enseña que el ingreso y los precios afectan conjuntamente a la demanda, por tanto, si aislamos el ingreso de la ecuación de la demanda no esperamos obtener un buen estimador para la elasticidad del precio. Sin embargo, y en situaciones más complicadas, no suele ser tan evidente averiguar cuáles son las variables a incorporar en una relación, lo que puede llegar a convertirse en un importante problema de especificación.

2. ***Inclusión de variables irrelevantes (redundantes).*** Caso contrario al anterior. Ahora, la hipótesis incluye variables que no deberían estar presentes en la ecuación. Este hecho tiene ciertas consecuencias sobre los procedimientos de inferencia aunque, en general, suelen ser menos graves que aquellas relacionadas con la exclusión de variables relevantes.

Existen contrastes para observar si un modelo adolece de variables omitidas. El ***test de la razón de verosimilitud para variables omitidas*** permite añadir un conjunto de variables a una ecuación existente y contrastar si constituyen una contribución significativa a la explicación de la variable dependiente. Este contraste tiene como hipótesis nula que el conjunto regresores adicionales no son conjuntamente significativos.

También existen contrastes para detectar si un modelo presenta variables redundantes. El ***test de la razón de verosimilitud para variables redundantes*** permite contrastar si un subconjunto de variables de una ecuación existente son conjuntamente significativas, o mejor dicho, si los coeficientes de determinadas variables del modelo van a tener valor cero, en cuyo caso esas variables pueden ser eliminadas de la ecuación del modelo.

El ***test de Wald para contrastar restricciones en los coeficientes de un modelo*** también puede utilizarse para detectar cuándo una variable es redundante. Basta comprobar cuándo puede considerarse cero su coeficiente de modo formal a través del test citado.

También es posible aplicar métodos de inclusión o exclusión automática de variables en el modelo. Partiendo de un conjunto inicial de k variables, se trata de no incluir nuevas variables irrelevantes en la definición del modelo ni omitir variables adecuadas. El método más rudimentario sería efectuar todas las regresiones posibles partiendo del conjunto más amplio de variables candidatas en el modelo y elegir la mejor con las variables que sean significativas. No obstante existen métodos automatizados que realizan esta tarea. El ***método de selección hacia adelante (método forward)*** permite partir de un conjunto mínimo de variables en la regresión e ir incluyendo variables adecuadas en el modelo de forma sucesiva comprobando la significatividad del nuevo coeficiente. El ***método de selección hacia atrás (método backward)*** parte de la regresión con todas las variables y va eliminando las no significativas por orden de significatividad hasta encontrar un modelo adecuado con todas sus variables lo suficientemente significativas. El ***método paso a paso (método stepwise)*** es un método de selección hacia adelante que comprueba en cada paso, no sólo la significatividad del nuevo coeficiente, sino también la de las variables incorporadas en los pasos anteriores.

Error de especificación en la forma funcional

Puede darse el caso de que las variables incluidas en un modelo sean las correctas pero la forma funcional lineal que las relaciona sea incorrecta. A veces, el contexto de modelo lineal es suficiente para manejar el problema, pero en ocasiones no puede sostenerse la linealidad y estamos ante problemas de no linealidad.

Una relación $Y = f(X_2, X_3)$ puede especificarse mediante la expresión $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$ o, como:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_2 X_2^2 + \gamma_3 X_3^2 + \delta(X_2 X_3) + u .$$

La segunda ecuación permite tanto una respuesta cuadrática a los regresores como un efecto de interacción. El efecto de interacción se basa en una nueva variable, el producto de los dos regresores. Por tanto, el efecto esperado de un cambio unitario en X_2 será $\beta_2 + 2\gamma_2 + \delta X_3$, dependiendo pues de β_2 y de los niveles de X_2 y X_3. Del mismo modo, el efecto esperado de un cambio unitario en X_3 dependerá tanto del nivel de X_2, como del de X_3. Cuando el error de especificación consiste en utilizar la primera ecuación en lugar de la segunda, aquél se corrige fácilmente añadiendo los términos X_2^2, X_3^2, y $(X_2 X_3)$. En otros casos, será necesaria una especificación intrínsicamente no lineal.

Para detectar problemas de especificación en la forma funcional suelen utilizarse los gráficos de los residuos, que, ante la presencia de no linealidades, normalmente presentan tendencias que indican su falta de aleatoriedad. Los altos grados de autocorrelación también son indicadores de la posible presencia de mala especificación funcional en el modelo. Asimismo, es útil realizar la representación de los gráficos de los residuos contra las variables explicativas y predichas con la finalidad de comprobar que son aleatorios. La falta de aleatoridad en estos gráficos puede indicar la presencia de un problema de no linealidad o de mala especificación funcional del modelo.

La solución para los problemas de mala especificación del modelo pasa por introducir variables ficticias o por la definición alternativa de la ecuación del modelo basándose en la tendencia observada en los gráficos residuales citados anteriormente.

Los errores de especificación debidos a la formulación incorrecta de la ecuación de un modelo de regresión lineal pueden contrastarse formalmente a través del ***contraste RESET de Ramsey***. Partiendo de que cualquier función puede ser aproximada por polinomios del orden adecuado, se puede introducir en el modelo de regresión términos con las potencias sucesivas y los productos cruzados de las variables explicativas. El contraste de Ramsey realiza un test para comprobar si los coeficientes de las potencias y productos cruzados incluidos en el modelo son cero, en cuyo caso se podría aceptar la forma funcional lineal del mismo. El contraste RESET de Ramsey sirve para detectar, además de una forma lineal incorrecta, cualquier error de omisión o la presencia de correlaciones entre variables explicativas y la perturbación.

EXOGENEIDAD Y REGRESORES ESTOCÁSTICOS

Un supuesto importante del modelo lineal es que ***los regresores no son estocásticos*** (no son aleatorios). Esta hipótesis, junto con las relativas al error o perturbación, implica la existencia de independencia entre dicha perturbación y las variables explicativas. Sin embargo hay modelos en los que la variable endógena depende de ella misma retardada o de variables cuya medición presenta algún error.

En estos casos aparecen dependencias entre la perturbación y las variables explicativas (***regresores exógenos***) que deben analizarse para no invalidar las propiedades que cumplen los estimadores por mínimos cuadrados ordinarios. El estimador MCO es sesgado en estos casos, pero según sea la dependencia puede ser consistente o inconsistente. También ocurre lo mismo en los modelos en que se cometen ***errores de medida en las variables*** provocados por inobservabilidad, mediciones repetidas, etc.

El método de las variables instrumentales

El método de las variables instrumentales (VI) permite obtener estimadores consistentes de los parámetros en situaciones en que MCO es inconsistente (omisión de variables relevantes, regresores exógenos, errores de medida o simultaneidad).

Supongamos que tenemos un vector Z de variables explicativas aleatorias. Si las variables Z son independientes de la perturbación u ($E(u|Z) = E(u) = 0$ y $E(uu'|Z = E(uu'))$, la estimación del modelo se realiza mediante MCO. Pero si las variables aleatorias Z no son independientes de u se usa el ***método de las variables instrumentales***, consistente en sustituir las variables problemáticas de Z por las denominadas instrumentos muy correlacionadas con ellas e incorrelacionadas con u. Después se aplica MCO.

Es necesario ser muy cuidadoso con la elección de cada instrumento, ya que no es evidente ver que está incorrelacionado con el error. Comprobar que el instrumento está correlacionado con la variable explicativa a la que sustituye ($C(Z,X) \neq 0$) puede hacerse a partir de una muestra. Basta plantear el siguiente modelo en forma reducida:

$$X = \pi_0 + \pi_1 Z + v$$

estimarlo por MCO y contrastar:

$$H_0: \pi_1 = 0 \quad versus \quad H_1: \pi_1 \neq 0$$

En el modelo de regresión múltiple con variable instrumental Z para X_2:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
$$E(\varepsilon) = 0, C(X_1, \varepsilon) = 0, C(X_2, \varepsilon) \neq 0$$
$$C(Z, \varepsilon) = 0$$

la forma reducida será:

$$X_2 = \pi_0 + \pi_1 X_1 + \pi_2 Z + \upsilon$$

Para que Z sea un instrumento válido será necesario que:

$$\pi_2 \neq 0$$

El estimador de mínimos cuadrados en dos etapas MC2E

Consideramos el modelo:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
$$C(X_1, \varepsilon) \neq 0$$

Supongamos que disponemos de dos posibles variables instrumentales Z_1 y Z_2 que cumplen (la C indica correlación):

$$C(Z_1, \varepsilon) = 0, C(Z_2, \varepsilon) = 0$$

El *Estimador de Mínimos Cuadrados en 2 Etapas* (MC2E) emplea como instrumento una combinación lineal de Z_1 y Z_2 y se obtiene de la siguiente forma:

1ª *Etapa*. Se estima por MCO la forma reducida del modelo:

$$X_1 = \pi_0 + \pi_1 Z_1 + \pi_2 Z_2 + \pi_3 X_2 + \upsilon$$

2ª *Etapa*. Se estiman α, β_1, y β_2 por MCO en la regresión de Y sobre $\hat{X}_1$ y X_2.

El contraste de Hausman

El contraste más usual para estudiar la exogeneidad de los regresores es el ***contraste de exogeneidad de Hausman***. Bajo la hipótesis nula se supone que el regresor no es estocástico, mientras que la alternativa establece lo contrario. El contraste compara los estimadores de los parámetros del modelo obtenidos por mínimos cuadrados ordinarios y por variables instrumentales. Bajo el supuesto de que el regresor no es estocástico (hipótesis nula), los estimadores obtenidos por ambos métodos son consistentes, mientras que bajo la hipótesis alternativa (regresor estocástico) sólo el estimador de variables instrumentales mantiene esta propiedad. Cuanto mayor sea la diferencia entre las estimaciones de los parámetros por ambos métodos, mayor evidencia existirá en contra de la hipótesis nula, es decir, mayor evidencia de presencia de regresores estocásticos.

El contraste de Hausman puede realizarse alternativamente mediante regresiones auxiliares. Se parte de un modelo con una sola variable explicativa cuyo carácter estocástico estamos estudiando $Y = \beta_0 + \beta_1 x + \varepsilon$. El contraste de Hausman es equivalente a contrastar la significatividad individual de la variable u en la regresión auxiliar $Y = \beta_0 + \beta_1 x + \delta u + \varepsilon$, donde u son los residuos de la regresión de la variable x sobre su variable instrumental utilizada z ($u_t = x_t - (\hat{\alpha}_0 + \hat{\alpha}_1 z_t)$).

ANÁLISIS DE LA INFLUENCIA

El análisis de la influencia tiene como finalidad conocer las observaciones muestrales que tienen una mayor influencia en la estimación del modelo y las observaciones atípicas o heterogéneas que no se ajustan al modelo

Influencia a priori (Leverage)

Para el estudio de la influencia de una observación en la estimación del modelo de regresión se debe tener en cuenta la siguiente ecuación:

$$\hat{y}_t = \hat{\alpha}_0 + \hat{\alpha}_1 x_{t1} + \cdots + \hat{\alpha}_k x_{tk} = \sum_{i=1}^{n} h_{ti} y_i \quad t = 1,\ldots,n$$

donde h_{ti} son unos pesos que en el modelo de regresión lineal simple (k=1) tienen la forma:

$$h_{ti} = \frac{1}{h} + \frac{(\bar{x}_t - \bar{x})(\bar{x}_i - \bar{x})}{\sum_{i=1}^{n}(\bar{x}_i - \bar{x})^2} \quad t,i = 1,\cdots,n$$

La ecuación anterior en forma matricial se expresa como:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\alpha} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\mathbf{Y}} = \mathbf{H}\vec{\mathbf{Y}}$$

siendo $H = X(\mathbf{X}^t\mathbf{X})^{-1}X^t$ la matriz de proyección ortogonal en el espacio generado por las variables regresoras. $H = \{h_{t,i}\}_{t,i=1^n}$ es una matriz cuadrada y simétrica.

De la ecuación de $\hat{\mathbf{Y}}$ se deduce que la predicción de una observación $\hat{y}_t$ es una combinación lineal de los valores de la variable respuesta $(\vec{Y} = (y_1,\ldots,y_n))$, con pesos h_{ti} que se obtienen a partir de los valores de la matriz diseño X. La influencia de la observación $(\vec{x}_i \ldots y_i)$ en el cálculo de $\hat{y}_t$ viene dado por:

- El valor de y_i
- El valor de h_{ti}

Por tanto, el valor de h_{ti} mide, al menos parcialmente, la influencia "a priori" de la observación i-ésima en el cálculo de $\hat{y}_t$. Su expresión viene dada por:

$$h_{ii} = \vec{x}_{i.}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{x}_{i.} \quad i = 1,2,\ldots,n$$

donde $\vec{x}_{i.}^t$ es la fila i-ésima de la matriz X (datos de la observación i-ésima).

En particular, en el modelo de regresión lineal simple (k = 1) se verifica:

$$h_{ii} = \frac{1}{n}\left(1 + \frac{(\bar{x}_i - \bar{x})^2}{s_x^2}\right) \quad i = 1,\ldots,n$$

En resumen, la influencia a priori de las observaciones viene dada por los elementos de la diagonal de H, h_{ii}, i = 1,2,...,n, el valor de h_{ii} mide la distancia al punto $\vec{x}_{i.}$ al centro $\bar{x}$, y se le denomina valor de *influencia a priori* (en inglés *leverage*). Observaciones con valor de influencia alto son observaciones que "a priori" influyen en el cálculo del modelo y observaciones con valor de influencia bajo "a priori" influyen poco.

Para saber si un h_{ii} es un valor grande o no se debe de tener en cuenta que si no hay filas repetidas en la matriz de diseño X se verifica que:

$$\frac{1}{n} \le h_{ii} \le 1 \quad i = 1,2,\ldots,n$$

$$\sum_{i=1}^{n} h_{ii} = traza(H) = k + 1$$

Por tanto $E^{(h_{ii})} = \dfrac{k+1}{n}$. Y se puede considerar que una observación tiene un valor de influencia grande si se verifica que $h_{tt} > 2\dfrac{k+1}{n}$.

Otro criterio se basa en calcular la varianza de los h_{ii}

$$s_h^2 = \frac{1}{n}\sum\left(h_{ii} - \frac{k+1}{n}\right)^2$$

y considerar que una observación tiene un valor de influencia grande si:

$$h_{ii} > E(h_{ii}) + 3s_h = \frac{k+1}{n} + 3s_h$$

El valor de influencia de las observaciones muestrales es un valor comprendido entre $1/n$ y 1, siendo los casos extremos:

$\bar{x}_{i.} = \bar{x}$, entonces $h_{ii} = 1/n$.

Considérese la muestra $\{(x^*, y_1), (x^*, y_2), \ldots, (x^*, y_{n-1}), (x_n, y_n)\}$ de un modelo de regresión lineal simple, entonces $h_{ii} = 1/(n-1)$ $i = 1, \ldots, n-1$, puntos en los que $x_i = x$ y $h_{nn} = 1$, el mayor valor que puede tomar. En este caso la recta de regresión pasa por los puntos $x^*, \bar{y}_{(n-1)}$ y (x_n, y_n), siendo $\bar{y}_{(n-1)} = 1/(n-1)\sum_{i=1}^{n-1} y_i$.

Unas pocas observaciones con valor de influencia a priori grande pueden producir multicolinealidad entre dos o más variables regresoras.

Distancia de Mahalanobis

Muchos paquetes estadísticos proporcionan la distancia de Mahalanobis de los puntos muestrales $\vec{x}_{i.} = (x_{i1}, x_{i2}, \ldots, x_{ik})$ $i = 1,2,\ldots,n$, al punto medio de la nube de las variables regresoras $\bar{x} = (\bar{x}_{.1}, \bar{x}_{.2}, \ldots, \bar{x}_{.k})$, donde $\bar{x}_{.j}$, $j = 1,2,\ldots,k$ es la media de los datos de la variable x_j. Esta distancia viene definida por la expresión $d_M^2(\vec{x}_{i.}; \bar{x}) = (\vec{x}_{i.} - \bar{x})\mathbf{S}^{-1}(\vec{x}_{i.} - \bar{x})'$ siendo S la matriz de varianzas-covarianzas del vector de variables $(x_1, x_2, \ldots, x_k)$.

La distancia de Mahalanobis es una distancia estadística que generaliza la distancia euclídea entre dos vectores en la que se tiene en cuenta la dispersión de las variables y su dependencia. Un valor alto de la distancia de Mahalanobis indica que el punto se aleja del centro de la nube y, por tanto, es una posible observación influyente a priori.

Criterio de los residuos eliminados

Si en el ajuste del modelo de regresión lineal no se utiliza la observación i-ésima, el vector de predicciones es:

$$\hat{\mathbf{Y}}_{(i)} = \mathbf{X}(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}'\hat{\mathbf{Y}}_{(i)}$$

donde el subíndice (i) indica que no se utiliza la observación i-ésima.

Se define el residuo eliminado $e_{(i)}$ como residuo obtenido utilizando la predicción calculada a partir de la muestra excepto la i-ésima observación, $\hat{y}_{i(i)}$. Esto es, $e_{(i)} = y_i - \hat{y}_{i(i)}$, $i = 1, \ldots, n$.

Teniendo en cuenta la siguiente relación entre los residuos ordinarios y los eliminados:

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}, \quad i = 1, \ldots, n$$

se puede deducir un nuevo criterio para distinguir a las observaciones influyentes a priori. Si la observación i-ésima influye mucho (h_{ii} es grande) los residuos ordinarios y los residuos eliminados son distintos, por el contrario, si el valor de influencia es pequeño ($h_{ii} \approx 0$) los dos residuos (ordinario y eliminado) son parecidos.

Influencia a posteriori: estadísticos de Cook y DFFITS

La identificación de las observaciones influyentes a posteriori es de mayor interés. Una observación influyente a posteriori es aquella $(\vec{x}_{i.}, y_i)$ cuya inclusión en el ajuste modifica sustancialmente la estimación del modelo. En este caso, se están considerando los datos de las variables regresoras y de la variable respuesta. El problema básico es determinar la influencia del dato $(\vec{x}_{i.}, y_i)$ en el ajuste del modelo de regresión lineal múltiple. O, equivalentemente, se desea realizar el siguiente contraste estadístico (C_I):

H_0 : el modelo ajustado con toda la muestra es igual al modelo ajustado con la muestra, excepto el dato $(\vec{x}_{i.}, y_i)$.

H_1 : el modelo ajustado con toda la muestra es distinto al modelo ajustado con la muestra, excepto el dato $(\vec{x}_{i.}, y_i)$.

Si la observación $(\vec{x}_{i.}, y_i)$ es influyente en el modelo de regresión, ésto se observa en:

- la estimación de los parámetros del modelo de regresión $(\vec{\alpha})$: α
- el vector de predicción de las observaciones: $\hat{\mathbf{Y}}$
- la predicción de la respuesta en el punto i-ésimo: $\hat{y}_i$

Los estadísticos para resolver el contraste C_I se basan en calcular la distancia entre las estimaciones de cualquiera de los tres valores anteriores cuando se utiliza toda la muestra, es decir, $\hat{\alpha} - \hat{\mathbf{Y}} - \hat{y}_i$ y las mismas estimaciones cuando se utiliza toda la muestra excepto el dato $(\vec{x}_{i.}, y_i)$, es decir, $\hat{\alpha}_{(i)} - \hat{\mathbf{Y}}_{(i)} - \hat{y}_{i(i)}$.

Las tres distancias llevan al mismo estadístico, el ***D-estadístico de Cook***, definido por:

$$D_{(i)} = \frac{\left(\hat{\alpha} - \hat{\alpha}_{(i)}\right)' \mathbf{X}'\mathbf{X}\left(\hat{\alpha} - \hat{\alpha}_{(i)}\right)}{(k+1)\hat{s}_R^2} = \frac{\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}\right)'\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}\right)}{(k+1)\hat{s}_R^2} =$$

$$\frac{\left(\hat{y}_i - \hat{y}_{i(i)}\right)^2}{(k+1)\hat{s}_R^2 h_{ii}} = \left(\frac{h_{ii}}{1-h_{ii}}\right)\frac{e_i^2}{(k+1)\hat{s}_R^2(1-h_{ii})} = \left(\frac{h_{ii}}{1-h_{ii}}\right)\frac{r_i^2}{k+1}$$

siendo r_i el i-ésimo residuo estandarizado y k el número de variables regresoras. Bajo la hipótesis nula, la observación i-ésima no es una observación influyente a posteriori, se verifica que:

$$D(i) \approx F_{k+1;n-(k+1)}$$

La ***familia de estadísticos DFFITS*** relacionados con el D-estadístico de Cook se definen como:

$$DFFITS(i) = \frac{\left(\hat{y}_i - \hat{y}_{i(i)}\right)^2}{\hat{s}_{R,(i)}\sqrt{h_{ii}}} = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}\frac{e_i}{\hat{s}_{R,(i)}\sqrt{1-h_{ii}}} = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} t_i$$

donde t_i es el residuo estudentizado.

Belsey, kuh y Welsch (1989) proponen utilizar como cota superior de este estadístico el valor $2\sqrt{\frac{k}{n}}$.

Esto es, la observación $(\vec{x}_{i.}, y_i)$ es influyente a posteriori si se cumple:

$$DFFITS(i) > 2\sqrt{\frac{k}{n}}$$

Ejercicio 3-1. Consideramos el modelo $y_t = \beta_0+\beta_1 x_1 + u_t$ y los siguientes datos para su ajuste:

y_t	6	3	1	1	1	4	6	16	25	36	49	64
x_t	−4	−3	−2	−1	1	2	3	4	5	6	7	8

Estimar el modelo por MCO y hallar el coeficiente de determinación corregido, así como el estadístico de Durbin-Watson. ¿Es conveniente reformular el modelo? En caso de respuesta positiva estimar el modelo alternativo más adecuado.

Tenemos que:

$$X' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -4 & -3 & -2 & -1 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{pmatrix}$$

$$Y' = (6 \;\; 3 \;\; 1 \;\; 1 \;\; 1 \;\; 4 \;\; 6 \;\; 16 \;\; 25 \;\; 36 \;\; 49 \;\; 64)$$

Los parámetros estimados serán:

$$\hat{B} = (X'X)^{-1}X'Y = \begin{pmatrix} 8{,}01 \\ 4{,}46 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

Por tanto, el modelo de regresión lineal ajustado será:

$$\hat{y} = 8{,}01 + 4{,}46x_1 + u$$

Para calcular el coeficiente de determinación utilizamos la expresión:

$$R^2 = \frac{SE}{ST} = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} = \frac{\hat{Y}'Y - T\bar{Y}^2}{\hat{Y}'Y - T\bar{Y}^2} = \frac{(X\hat{\beta})'Y - T\bar{Y}^2}{\hat{Y}'Y - T\bar{Y}^2} = 0{,}7$$

Este resultado significa que el 70% de la variación en la variable dependiente se explica con nuestro modelo de regresión lineal.

El coeficiente de determinación corregido se calcula como sigue:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{T-1}{T-k-1} = 0{,}67$$

Para contrastar la significatividad conjunta de los parámetros del modelo, es decir, $\beta_0 = \beta_1 = 0$, utilizamos el estadístico:

$$F_{k,T-k-1} = \frac{SE/k}{SR/(T-k-1)} = \frac{R^2/k}{(1-R^2)/(T-k-1)} = \frac{0{,}7/1}{(1-0{,}7)/(12-1-1)} = 23{,}3$$

Como el p-valor $P(F_{1,10} > 23{,}3) = 0{,}0007 < 0{,}005$ se concluye la significatividad conjunta del modelo (rechazo de $\beta_0 = \beta_1 = 0$).

Para analizar la significatividad individual de los parámetros estimados calculamos:

$$T_0 = \frac{\hat{\beta}_0}{\hat{\sigma}\sqrt{a_{00}}} = 1{,}97 \qquad T_1 = \frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{a_{11}}} = 4{,}84$$

No olvidemos que a_{ii} son los elementos de la diagonal de $(X'X)^{-1}$, y que:

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{T-k-1} = \frac{(Y-X\hat{B})'(Y-X\hat{B})}{12-1-1} = \frac{Y'Y-(X\hat{B})'Y}{10} = 150{,}1$$

Los p-valores asociados a los T_i, según una distribución T de Student con 10 grados de libertad son $P(T > 1{,}97) = 0{,}0766$ y $P(T > 4{,}84) = 0{,}0007$.

Según los p-valores, la constante no es significativamente distinta de cero al 95%. Para asegurar la significatividad de los dos parámetros es necesario rebajar el nivel de confianza al 93% como mínimo.

Como el coeficiente de determinación ajustado es bajo y hay un parámetro no muy significativo, puede ocurrir que existan problemas en el modelo. A continuación calcularemos el estadístico de Durbin-Watson para analizar la autocorrelación.

Para ello calculamos previamente los residuos estimados $\hat{u} = Y - X\hat{B}$ obteniéndose:

$\hat{u}' = (8{,}35,\ 1{,}89,\ -2{,}55,\ -11{,}46,\ -12{,}92,\ -15{,}37,\ -9{,}83,\ -5{,}29,\ 1{,}25,\ 9{,}79,\ 20{,}34)$

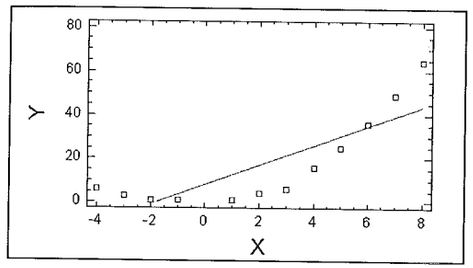
El estadístico de Durbin-Watson se calcula como sigue:

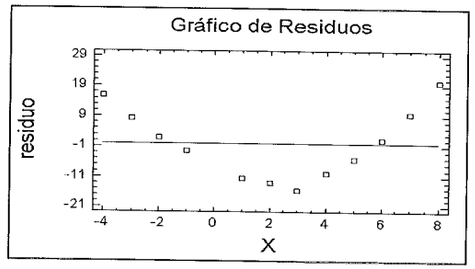
$$d = \frac{\sum_{t=1}^{T}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T}\hat{u}_t^2} = 0{,}32$$

En las tablas de Durbin-Watson para 12 observaciones y $\alpha = 0{,}05$ se tiene $d_L = 0{,}971$ y $d_U = 1{,}331$.

Está claro que existe autocorrelación positiva, ya que $d = 0{,}32 < d_L = 0{,}971$.

Para intentar arreglar el problema de la autocorrelación representamos las variables gráficamente para ver si siguen una tendencia distinta de la lineal. En la gráfica se observa que se ajusta mejor una parábola que una recta a la nube de puntos. Por otra parte, si representamos los residuos contra los valores de la variable independiente obtenemos también tendencia cuadrática.





De las figuras anteriores se deduce que hemos cometido un error de especificación en el modelo, siendo más adecuado el modelo cuadrático. Por tanto se propone el modelo:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + u_t$$

Si estimamos este nuevo modelo, tenemos:

Parámetro	Estimación	T	P-Valor
CONSTANTE	-1,78169	-2,40505	0,0396
X	1,03343	5,00579	0,0007
X^2	0,88252	21,6641	0,0000

R-cuadrado = 99,4384 %
R-cuadrado (ajustado) = 99,3135 %
Estadístico de Durbin-Watson = 1,21575
Valor de la *F* de Fisher-Snedecor = 796,72
p-valor para la *F* = 0,0000

El nuevo ajuste del modelo es mejor que el anterior ya que los coeficientes de determinación y el estadístico de Durbin-Watson han mejorado. Además, todas las variables son significativas individualmente a partir de un 96% y en conjunto al 100% (p-valor de la F óptimo).

Ejercicio 3-2. Consideramos el modelo keynesiano del consumo C en función de la renta Y definido como:

$$C_t = \beta_0 + \beta_1 Y_t + u_t$$

Dados los datos siguientes:

Año	*C*	*Y*	*Año*	*C*	*Y*
1954	25268	27889	1970	51713	57850
1955	26445	28816	1971	53794	60615
1956	28092	31229	1972	57725	65061
1957	28768	32273	1973	61827	70050
1958	29780	32621	1974	64386	72818
1959	30034	31515	1975	65250	73565
1960	28444	31173	1976	67481	74931
1961	31277	34762	1977	68359	74841
1962	33826	37736	1978	68507	75990
1963	37338	42059	1979	68713	75517
1964	38600	43742	1980	69081	74572
1965	40845	46419	1981	67964	73751
1966	43153	49263	1982	67913	73593
1967	45197	49774	1983	67861	72530
1968	47295	52140	1984	66851	71539
1969	50115	55260			

Estimar el modelo con los datos del periodo 1955-1984 y analizar la propensión marginal al consumo (coeficiente de la renta). Analizar la autocorrelación del modelo, y, en caso de existir problemas, considerar el modelo alternativo del consumo de Brown formulado como:

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 C_{t-1} + u_t$$

Estimar este nuevo modelo analizando también su autocorrelación y heteroscedasticidad.

Los parámetros estimados serán:

$$\hat{B} = (X'X)^{-1}X'Y = \begin{pmatrix} 30 & 1666005 \\ 1666005 & 101384915993 \end{pmatrix}^{-1} \begin{pmatrix} 15506634 \\ 91762844616 \end{pmatrix} =$$

$$\begin{pmatrix} 0{,}3811823844 & -0{,}0000062638 \\ -0{,}0000062638 & 1{,}1279263(10)^{-10} \end{pmatrix}^{-1} \begin{pmatrix} 15506634 \\ 91762844616 \end{pmatrix} = \begin{pmatrix} -485{,}221 \\ 0{,}913 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

Por tanto, el modelo de regresión lineal ajustado será:

$$C = -485{,}221 + 0{,}913Y + u$$

La suma de los cuadrados de los residuos viene dada por:

$$\hat{u}'\hat{u} = y'y - \hat{\beta}'X'y = 32823077$$

La estimación de la varianza residual viene dada por:

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{T-k-1} = \frac{32823077}{30-1-1} = \frac{Y'Y - (X\hat{B})'Y}{28} = 1172252{,}8$$

Las desviaciones típicas de los estimadores vienen dadas por:

$$\hat{\sigma}(\hat{\beta}_0) = \sqrt{1172252{,}8(0{,}3811823844)} = 668{,}556$$

$$\hat{\sigma}(\hat{\beta}_1) = \sqrt{1172252{,}8(1{,}1272252{,}8)10^{-10}} = 0{,}011$$

Para analizar la propensión marginal al consumo comprobamos la significatividad individual del parámetro $\hat{\beta}_1$ a partir del estadístico:

$$T_1 = \frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{a_{11}}} = \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)} = \frac{0{,}913}{0{,}011} = 83$$

Además, el estadístico anterior se distribuye según una t de Student con 28 grados de libertad cuyo valor crítico al 95% es $t_{0{,}05;28} = 1{,}701$. Entonces, para realizar el contraste:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 > 0$$

se observa que $1{,}701 < 83$, por lo que rechazamos la hipótesis nula, aceptando que la propensión marginal al consumo es positiva.

También podemos contrastar si la propensión marginal al consumo es unitaria, es decir:

$$H_0: \beta_1 = 1$$
$$H_1: \beta_1 < 1$$

Para ello consideramos el estadístico:

$$T'_1 = \frac{\hat{\beta}_1 - 1}{\hat{\sigma}\sqrt{a_{11}}} = \frac{\hat{\beta}_1 - 1}{\hat{\sigma}(\hat{\beta}_1)} = \frac{0{,}913 - 1}{0{,}011} = -7{,}9099$$

Se observa que -7,909 < -1,701, por lo que rechazamos la hipótesis nula, aceptando que la propensión marginal al consumo es menor que la unidad.

No obstante, todos los cálculos anteriores los hemos realizado sin contrastar la validez del modelo. Vamos ahora a analizar si nuestro modelo presenta problemas de autocorrelación analizando el estadístico de Durbin-Watson *d*. Para ello sabemos que:

$$d = 2(1 - \hat{\rho})$$

Además:

$$\hat{\rho} = \frac{\sum_{t=1}^{T} \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^{T} \hat{u}_t^2} = 0{,}72 \Rightarrow d = 2(1 - 0{,}72) = 0{,}55$$

En la tabla del estadístico de Durbin-Watson para $T = 30$, $k = 2$ y $\alpha = 0{,}05$ se tiene que $d_L = 1{,}352$ y $d_U = 1{,}489$. Como $d = 0{,}55 < d_L = 1{,}352$ se acepta la existencia de autocorrelación positiva de primer orden.

La presencia de autocorrelación nos lleva a considerar como alternativa posible a la modelización del consumo el modelo Brown. Realizaremos su estimación como sigue:

$$\hat{B} = (X'X)^{-1} X'Y = \begin{pmatrix} 521{,}510 \\ 0{,}532 \\ 0{,}413 \end{pmatrix}$$

Por tanto tenemos:

$$\hat{\beta}_0 = 521{,}510, \quad \hat{\beta}_1 = 0{,}532, \quad \hat{\beta}_2 = 0{,}413$$

Por tanto, el modelo de regresión lineal múltiple ajustado será:

$$C_t = 512{,}510 + 0{,}532 Y_t + 0{,}413 C_{t-1} + u_t$$

El estimador de la varianza vendrá dado por la expresión:

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{T-k-1} = \frac{Y'Y - (X\hat{B})'Y}{30-2-1} = \frac{6704709}{27} = 248322{,}45$$

No olvidemos que las matrices Y, X y $\hat{B}$ son ya conocidas.

Para calcular el coeficiente de determinación utilizamos la expresión:

$$R^2 = \frac{SE}{ST} = \frac{\sum_{t=1}^{T} (\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2} = \frac{\hat{Y}'Y - T\bar{Y}^2}{\hat{Y}'Y - T\bar{Y}^2} = \frac{(X\hat{\beta})'Y - T\bar{Y}^2}{\hat{Y}'Y - T\bar{Y}^2} = 0{,}9991$$

Este resultado significa que el 99,91% de la variación en la variable dependiente se explica con nuestro modelo de regresión lineal.

El coeficiente de determinación corregido se calcula como sigue:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{T-1}{T-k-1} = 0{,}999$$

Para contrastar la significatividad conjunta de los parámetros del modelo, es decir, $\beta_0 = \beta_1 = \beta_2 = 0$, utilizamos el estadístico:

$$F_{k,T-k-1} = \frac{R^2/k}{(1-R^2)/(T-k-1)} = \frac{0{,}9991/2}{(1-0{,}9991)/(30-2-1)} = 14986{,}5$$

Como el p-valor $P(F_{2,27} > 14986{,}5) < 0{,}005$ se concluye la significatividad conjunta del modelo. También puede observarse en las tablas de la F que $F_{0{,}05;2;27} = 3{,}36 < 14986{,}5$, lo que nos lleva a la aceptación de la significatividad conjunta del modelo (rechazo de $\beta_0 = \beta_1 = \beta_2 = 0$).

Para analizar la significatividad individual de los parámetros estimados calculamos:

$$T_0 = \frac{\hat{\beta}_0}{\hat{\sigma}\sqrt{a_{00}}} = \frac{521{,}51}{322{,}984} = 1{,}61 \qquad T_1 = \frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{a_{11}}} = \frac{0{,}532}{0{,}037} = 14{,}37$$

$$T_2 = \frac{\hat{\beta}_2}{\hat{\sigma}\sqrt{a_{22}}} = \frac{0{,}413}{0{,}040} = 10{,}325$$

Los p-valores asociados a los T_i, según una distribución T de Student con 27 grados de libertad son P(T > 1,61) = 0,0197, P(T > 14,37) = 5,07(10) $^{-14}$ y P(T > 10,325) = 3,23(10) $^{-11}$.

Todos los p-valores son menores que 0,05, con lo que todos los parámetros (incluida la propensión marginal al consumo) son significativamente distintos de cero al 95%. Se observa que ahora la propensión marginal al consumo se estima en 0,532.

Además, como:

$$T_1 = \frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{a_{11}}} = \frac{0{,}532}{0{,}037} = 14{,}37 > t_{0{,}05;27} = 1{,}704$$

se acepta la hipótesis de que la propensión marginal al consumo es positiva (contraste de $\hat{\beta}_1 = 0$ contra $\hat{\beta}_1 > 0$).

Por otra parte, como:

$$T'_1 = \frac{\hat{\beta}_1 - 1}{\hat{\sigma}\sqrt{a_{11}}} = \frac{0{,}532 - 1}{0{,}037} = -12{,}64 < -1{,}704$$

se acepta la hipótesis de que la propensión marginal al consumo es menor que la unidad (contraste de $\hat{\beta}_1 = 1$ contra $\hat{\beta}_1 < 1$).

A continuación, calculamos el estadístico de Durbin-Watson:

$$d = \frac{\sum_{t=1}^{T}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T}\hat{u}_t^2} = 2{,}104$$

Se obtiene un valor de d muy próximo a 2, pero como el modelo tiene como regresor el regresando desfasado, no es adecuado utilizar el estadístico de Durbin-Watson para analizar la autocorrelación. En su lugar utilizaremos el contraste h de Durbin, cuyo estadístico se define como:

$$h = \hat{\rho}\sqrt{\frac{T}{1 - T(\hat{V}(\hat{\beta}_i))}} = \hat{\rho}\sqrt{\frac{T}{1 - T(\hat{\sigma}(\hat{\beta}_i))^2}}$$

Tenemos:

$$d = 2(1-\hat{\rho}) \Rightarrow \hat{\rho} = 1 - \frac{d}{2} = -0{,}052 \Rightarrow h = \hat{\rho}\sqrt{\frac{T}{1 - T(\hat{\sigma}(\hat{\beta}_i))^2}} = (-0{,}052)\sqrt{\frac{30}{1 - 30(0{,}04)^2}} = -0{,}292$$

Como |-0,292| < 1,96 se acepta la hipótesis nula de no autocorrelación para un nivel de significación de 0,05. Vemos entonces que el modelo de consumo de Brown está libre de autocorrelación.

A continuación, analizaremos la heteroscedasticidad usando el contraste de W de White. Para ello realizamos una regresión MCO de los cuadrados de los residuos sobre todas las variables independientes del modelo, sus cuadrados y sus productos cruzados. Obtenemos:

$$\hat{u}_t^2 = -43991{,}86 - 169{,}11Y_t + 208{,}29C_{t-1} + 0{,}0012Y_t^2 - 0{,}0014C_{t-1}^2 - 0{,}002Y_tC_{t-1}$$

La homoscedasticidad se aceptará si el contraste de la F permite aceptar la nulidad conjunta de todos los parámetros salvo la constante.

$$F_{k,T-k-1} = \frac{R^2/k}{(1-R^2)/(T-k-1)} = \frac{0{,}164/2}{(1-0{,}164)/(30-2-1)} = 2{,}648$$

Puede observarse en las tablas de la F que $F_{0{,}05;2;27} = 3{,}36 > 2{,}648$, lo que nos lleva al rechazo de la significatividad conjunta del modelo (aceptación de $\beta_0=\beta_1=\beta_2=0$). Por tanto el modelo de Brown está libre de heteroscedasticidad.

Una forma alternativa muy sencilla de realizar el contraste de White es utilizar una chi-cuadrado en vez del contraste de la F. Para ello, sabemos que bajo la hipótesis nula de homoscedasticidad, el estadístico $T(R^2)$ se distribuye según una chi-cuadrado con $p(p-1)/2$ -1 grados de libertad siendo p el número de regresores del modelo original incluyendo el término independiente (en nuestro caso $p = 4$). Luego tenemos:

$$T(R^2) \rightarrow \chi^2_5$$

En nuestro caso $T(R^2) = 30(0{,}164) = 4{,}92$. El valor tabular al 95% para una chi-cuadrado con 5 grados de libertad es 11,07 > 4,92, con lo que se acepta la hipótesis nula de homoscedasticidad.

Podemos concluir que el modelo de Brown es un buen modelo para ajustar el consumo en función de la renta disponible para los datos dados.

Ejercicio 3-3. Consideramos el modelo $Y_t = \beta_0 + \beta_1 X_t + u_t$ ***en el que se conoce que los residuos siguen la siguiente estructura AR(1)*** $u_t = 0{,}5u_{t-1} + e_t \quad e_t \to N(0,\sigma^2)$. ***El modelo se estima para los siguientes datos:***

Y_t	22	26	32	31	40	46	46	50
X_t	4	6	10	12	13	16	20	22

A la vista de esta información realizar una estimación correcta del modelo.

Estamos ante un modelo con autocorrelación en el que los residuos siguen un proceso autorregresivo de orden uno. Como además conocemos el valor de $\rho = 0{,}5$, podemos intentar la ***estimación por el método de mínimos cuadrados generalizados MCG***.

Las estimaciones vienen dadas por $\hat{B}_{MCG} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$ siendo:

$$\Omega^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & 0 & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Según los datos de nuestro problema:

$$\Omega^{-1} = \frac{1}{1-0{,}5^2}\begin{bmatrix} 1 & -0{,}5 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0{,}5 & 1{,}25 & -0{,}5 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0{,}5 & 1{,}25 & -0{,}5 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0{,}5 & 1{,}25 & -0{,}5 & 0 & .0 & 0 \\ 0 & 0 & 0 & -0{,}5 & 1{,}25 & -0{,}5 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0{,}5 & 1{,}25 & -0{,}5 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0{,}5 & 1{,}25 & -0{,}5 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0{,}5 & 1 \end{bmatrix}$$

Luego tenemos:

$$\hat{B}_{MCG} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y = \begin{pmatrix} 17.19 \\ 1{,}49 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

La matriz de varianzas covarianzas del estimador es:

$$\Sigma_{\hat{\beta}} = \sigma^2(X'\Omega^{-1}X)^{-1} = \sigma^2\begin{pmatrix} 1{,}2095 & -0{,}0705 \\ -0{,}0705 & 0{,}00546 \end{pmatrix}$$

El modelo correctamente estimado será:

$$Y_t = 17{,}19 + 1{,}49X_t + u_t$$

Alternativamente, puede obtenerse el ***estimador MCG mediante el método de Price Winsten*** estimando el modelo transformado:

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_{2t} - \rho X_{2t-1}) + \ldots + \beta_k(X_{kt} - \rho X_{kt-1})$$

para $t = 2,3,\ldots,T$ por MCO. Para $t = 1$ se realiza la transformación siguiente:

$$\sqrt{1-\rho^2}\,Y_1 = \sqrt{1-\rho^2}\,(\beta_1 + \beta_2 X_{21} + \cdots + \beta_k X_{k1} + u_1)$$

Si representamos el modelo transformado como $M_t = \beta'_0 Z_{0t} + \beta'_1 Z_{1t} + u_t$ Los datos transformados serán los siguientes:

	M_t	Z_{0t}	Z_{1t}
t	$\sqrt{1-\rho^2}\,Y_1$ (para $t = 1$) ó $Y_t - \rho Y_{t-1}$ (para $t = 2,..,8$)	$\sqrt{1-\rho^2}$ (para $t = 1$) ó $1-\rho$ (para $t = 2,..,8$)	$\sqrt{1-\rho^2}\,X_1$ (para $t = 1$) ó $Y_t - \rho X_{t-1}$ (para $t = 2,..,8$)
1	19,05255888	0,866025404	3,46410162
2	15	0,5	4
3	19	0,5	7
4	15	0,5	7
5	24,5	0,5	7
6	26	0,5	9,5
7	23	0,5	12
8	27	0,5	12

Ahora, las matrices X e Y para el ajuste MCO (sin constante) son las siguientes:

$$X = \begin{pmatrix} 0{,}86602504 & 3{,}4610 \\ 0{,}5 & 4 \\ 0{,}5 & 7 \\ 0{,}5 & 7 \\ 0{,}5 & 7 \\ 0{,}5 & 9{,}5 \\ 0{,}5 & 12 \\ 0{,}5 & 12 \end{pmatrix} \qquad Y = \begin{pmatrix} 19{,}05255888 \\ 15 \\ 19 \\ 15 \\ 24{,}5 \\ 26 \\ 23 \\ 27 \end{pmatrix}$$

$$\hat{B}_{MCG} = (X'X)^{-1}X'Y = \begin{pmatrix} 17.19 \\ 1{,}49 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

Se observa que los resultados coinciden al utilizar los dos métodos. Además, este último ajuste por MCO presenta características óptimas (significatividad fuerte de todos los parámetros, tanto individual como conjunta y buen valor del estadístico de Durbin Watson.

Parámetro	Estimación	Desviación típica	T	P-Valor
Z0	17,1935	4,34546	3,95665	0,0075
Z1	1,49663	0,292109	5,12353	0,0022

```
R-cuadrado = 98,1054 porcentaje
R-cuadrado (ajustado para g.l.) = 97,7897 porcentaje
Desviación típica del error = 3,42186
Estadístico de Durbin-Watson = 2,49141
F = 155,35 (p-valor = 0,0000).
```

Ejercicio 3-4. Tratamos de explicar las importaciones mediante el consumo y la inversión disponiendo de los siguientes datos:

IMP	*PIB*	*CONS*	*INV*	*IMP*	*PIB*	*CONS*	*INV*	*IMP*	*PIB*	*CONS*	*INV*
159	1493	1081	42	164	1612	1148	41	190	1715	1232	31
191	1755	1269	31	188	1808	1321	11	204	1907	1377	22
227	2021	1460	21	265	2124	1541	56	281	2261	1623	50
276	2319	164	51	263	2390	1676	7	311	2580	1768	56
333	2698	186	39	370	2884	1997	31	433	3045	2139	46
490	3234	2238	70	503	3368	2320	12	566	3539	2429	45

Ajustar de forma razonable el modelo $IMP = \beta_0 + \beta_1 PIB + \beta_2 CONS + \beta_3 INV + u$ ***analizando los posibles problemas.***

Si ajustamos el modelo por MCO se obtienen los resultados siguientes:

Variable dependiente: IMPOR

Parámetro	Estimación	Desviación estándar	T	P-Valor
CONSTANTE	-197,251	41,2525	-4,7815	0,0003
PIB	0,032204	0,186884	0,1723	0,8656
CONSUMO	0,242747	0,285361	0,8506	0,4093
INVERSIÓN	0,414199	0,32226	1,2853	0,2195

```
R-cuadrado = 97,3043 porcentaje
R-cuadrado (ajustado para g.l.) = 96,7267 porcentaje
Error estándar de est. = 22,5817
F = 168,45 (p-valor = 0,0000)
Estadístico de Durbin-Watson = 0,240325
```

Se observa la no significatividad de ninguna variable explicativa y la gran significatividad del modelo en su conjunto (p-valor de la F igual cero). Esto nos lleva a sospechar la presencia de multicolinealidad. También se observa autocorrelación residual (p-valor de Durban-Watson problemático).

Además, si ajustamos el modelo sin la última observación para cada variable, se obtienen estimadores muy distintos de los anteriores, lo que confirma la multicolinealidad. El resultado de este ajuste se presenta a continuación.

Variable dependiente: IMPOR

Parámetro	Estimación	Error estándar	T	P-Valor
CONSTANTE	-183,916	35,7713	-5,14143	0,0002
PIB	-0,01069	0,161129	-0,06637	0,9481
CONSUMO	0,29411	0,245478	1,19811	0,2523
INVERSIÓN	0,40099	0,276269	1,45145	0,1704

```
R-cuadrado = 97,4415 porcentaje
R-cuadrado (ajustado para g.l.) = 96,8511 porcentaje
Error estándar de est. = 19,3553
F = 165,04 (p-valor = 0,0000)
Estadístico de Durbin-Watson = 0,301764
```

Para estimar el modelo con multicolinealidad utilizaremos el procedimiento de *Regresión en cadena* tomando como valor del parámetro de cadena $c = 0{,}06$ (es el número entre 0,001 y 0,1 que mejor ajuste proporciona para el modelo). La regresión en cadena ofrece como estimadores de los parámetros $(X'X+cI)^{-1}\ X'Y$ siendo c la constante adecuada. La matriz de varianzas covarianzas toma la forma $\sigma^2(X'X+cI)^2X'X(X'X+cI)^{-1}$. En la práctica suele tomarse c como un valor entre 0,01 y 0,1 que hace que el ajuste sea bueno en cuanto a R^2 y significatividad individual y conjunta. El resultado para nuestros datos con c = 0,06 es el siguiente:

$$\hat{B} = \left(X'X + 0{,}06I^2\right)^{-1} X'Y = \begin{pmatrix} -172{,}54 \\ 0{,}091620 \\ 0{,}143353 \\ 0{,}428728 \end{pmatrix}$$

$$R^2 = \frac{SE}{ST} = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} = \frac{\hat{Y}'Y - T\bar{Y}^2}{\hat{Y}'Y - T\bar{Y}^2} = \frac{(X\hat{\beta})'Y - T\bar{Y}^2}{\hat{Y}'Y - T\bar{Y}^2} = 0{,}945078$$

Se obtienen entonces los siguientes resultados para la estimación del modelo ya sin multicolinealidad:

Parámetro	Estimación
CONSTANTE	-172,54
PIB	0,091620
CONSUMO	0,143353
INVERSION	0,428728

R-cuadrado = 94,5078 porcentaje
R-cuadrado (adaptado para g.l.) = 93,3309 porcentaje
Error estándar de la est. = 23,0059

T =18 es el número de observaciones, I es la matriz identidad y las matrices X e Y inmersas en los cálculos de este problema son las siguientes:

$$X = \begin{pmatrix} 1 & 1493 & 1081 & 42 \\ 1 & 1755 & 1269 & 31 \\ 1 & 2021 & 1460 & 21 \\ 1 & 2319 & 1643 & 51 \\ 1 & 2698 & 1866 & 39 \\ 1 & 3234 & 2238 & 70 \\ 1 & 1612 & 1148 & 41 \\ 1 & 1808 & 1321 & 11 \\ 1 & 2124 & 1541 & 56 \\ 1 & 2390 & 1676 & 7 \\ 1 & 2884 & 1997 & 31 \\ 1 & 3368 & 2320 & 12 \\ 1 & 1715 & 1232 & 31 \\ 1 & 1907 & 1377 & 22 \\ 1 & 2261 & 1623 & 50 \\ 1 & 2580 & 1768 & 56 \\ 1 & 3045 & 2139 & 46 \\ 1 & 3539 & 2429 & 45 \end{pmatrix} \qquad Y = \begin{pmatrix} 159 \\ 191 \\ 227 \\ 276 \\ 333 \\ 490 \\ 164 \\ 188 \\ 265 \\ 263 \\ 370 \\ 503 \\ 190 \\ 204 \\ 281 \\ 311 \\ 433 \\ 566 \end{pmatrix}$$

Ejercicio 3-5. Se trata de explicar la cantidad demandada Q1 del bien 1 en función de su precio P1, del nivel de renta disponible YD y de los precios P2 y P3 de dos bienes alternativos mediante el modelo:

$$Q1 = b_1 P1^{b_2} P2^{b_3} P3^{b_4} YD^{b_5}$$

Se dispone de los siguientes datos:

P1	*P2*	*P3*	*YD*	*Q1*	*P1*	*P2*	*P3*	*YD*	*Q1*
10000	*10000*	*10000*	*10000*	*10000*	*10242*	*8464*	*8585*	*10778*	*9374*
8915	*9787*	*9176*	*11113*	*11932*	*9705*	*9483*	*9983*	*11696*	*9894*
10526	*9045*	*10358*	*11853*	*9958*	*10282*	*11166*	*9807*	*12716*	*11733*
11335	*9420*	*10466*	*13331*	*9987*	*11290*	*9466*	*10010*	*13474*	*10095*
11233	*9418*	*10331*	*15368*	*11509*	*11416*	*8685*	*12901*	*15544*	*9566*
10293	*8434*	*11965*	*16132*	*10622*	*9777*	*7617*	*11830*	*16444*	*10781*
9877	*8442*	*10930*	*16693*	*13444*	*11740*	*8918*	*14591*	*18173*	*10131*
11153	*7987*	*13144*	*18800*	*10413*	*11230*	*8037*	*11058*	*18874*	*11597*
11201	*8064*	*13540*	*18913*	*10513*	*10971*	*8511*	*13316*	*19280*	*10564*
12083	*5962*	*14590*	*20427*	*9427*	*9640*	*7168*	*15322*	*20878*	*12541*
10818	*7201*	*14677*	*21662*	*11199*	*10930*	*5104*	*14025*	*21780*	*9301*
11203	*8234*	*13761*	*22211*	*11714*	*11742*	*7696*	*13023*	*22231*	*10845*
11465	*7589*	*15285*	*23113*	*14097*	*9908*	*7997*	*14246*	*23119*	*13655*
11744	*7224*	*14257*	*23425*	*13353*	*11600*	*6547*	*16598*	*24195*	*9382*
12485	*6636*	*17058*	*24216*	*9422*	*12269*	*7433*	*18204*	*25495*	*11734*

Ajustar de forma razonable el modelo analizando los posibles problemas.

El primer paso para trabajar con este modelo es linealizarlo aplicando logaritmos, con lo que se transformará en el modelo:

$$Log(Q1) = Log(b_1) + b_2 * Log(P1) + b_3 * Log(P2) + b_4 * Log(P3) + b5 * Log(YD) + u$$

Al ajustar el modelo obtenemos los siguientes resultados:

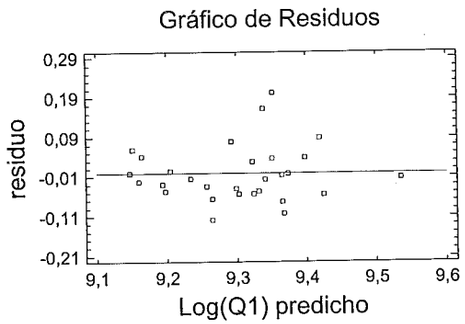
Variable dependiente: Log(Q1)

Parámetro	Estimación	Desviación típica	T	P-Valor
CONSTANTE	8,60848	2,408390	3,57438	0,0015
Log(P1)	-0,89012	0,215719	-4,12630	0,0004
Log(P2)	0,554186	0,130026	4,26212	0,0003
Log(P3)	-0,349092	0,183114	-1,90642	0,0682
Log(YD)	0,743952	0,135187	5,50313	0,0000

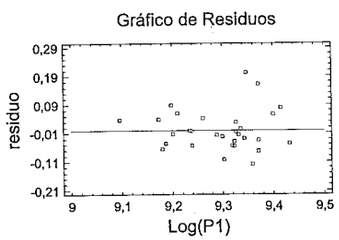
R-cuadrado = 66,4287 porcentaje
R-cuadrado (ajustado para g.l.) = 61,0573 porcentaje
Desviación típica del error σ = 0,0761884
F = 12,37 (p-valor = 0,0000)
Estadístico de Durbin-Watson = 2,47105

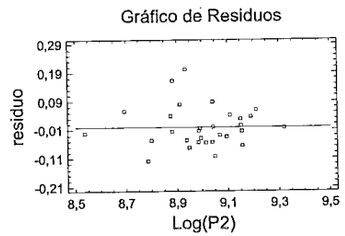
Inicialmente el ajuste parece correcto, pues hay significatividad al menos al 93% (p-valores de la T menores que 0,07) de todos los parámetros y también hay fuerte significatividad global (p-valor de la F nulo). Por otra parte no hay autocorrelación, ya que el p-valor del estadístico de Durban-Watson es mayor que 0,05. Quizás el R^2 sea algo bajo, pero no drásticamente.

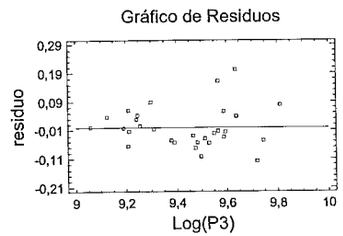
Pero el problema aparece al graficar los residuos contra los valores predichos, que muestra una estructura no demasiado aleatoria de sus puntos (este hecho nos lleva a sospechar la presencia de heteroscedasticidad) tal y como se observa en la figura siguiente:

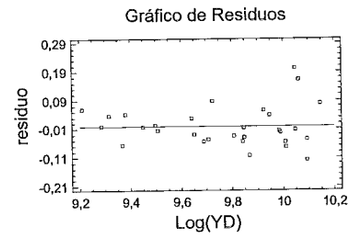


Para detectar qué variables son las responsables de la posible heteroscedasticidad realizamos los gráficos de residuos contra las cuatro variables explicativas. Se obtienen las figuras siguientes:









Observándose que la menos aleatoria es la relativa a *YD*, ya que aumenta la dispersión del error al ir de izquierda a derecha y presenta un ajuste bueno a una recta paralela al eje *X*; por tanto no tiene estructura aleatoria.

La variable *YD* es la candidata a provocar los problemas de heteroscedasticidad, pero para estar más seguros utilizaremos un contraste formal, por ejemplo el de Glesjer. Para ello realizamos una regresión múltiple de *Abs(RESIDUOS)* contra *Log(YD)*. Se obtienen los resultados que se presentan a continuación:

```
Variable dependiente: Abs(RESIDUOS)
-----------------------------------------------------------------------------
Parámetro              Estimación   Desviación típica     T            P-Valor
-----------------------------------------------------------------------------
CONSTANTE             -0,610484      0,2773170        -2,2014         0,0361
Log(YD)                0,067936      0,0283841         2,3934         0,0236
-----------------------------------------------------------------------------

R-cuadrado = 16,9847 porcentaje
R-cuadrado (ajustado para g.l.) = 14,0199 porcentaje
Error estándar de est. = 0,0424655
F = 5,73 (p-valor = 0,0236)
Estadístico de Durbin-Watson = 2,1298
```

Todos los coeficientes resultan significativos al 95% (p-valores menores que 0,05) y el modelo global también es significativo (p-valor de la F menor que 0,5). Por tanto, tenemos que rechazar la nulidad de cualquier coeficiente incluida la constante. Como, según el contraste de Glesjer, la hipótesis de homoscedasticidad es equivalente a la hipótesis de pendiente nula en este modelo, hay heteroscedasticidad según *YD*.

Ahora resta la tarea de estimar el modelo inicial con heteroscedasticidad. Para ello observamos en la figura de *YD* contra los residuos, que la dispersión residual es creciente con *YD*, lo que nos permite suponer como modelo para varianza residual $\sigma_i^2 = \sigma^2(log(YD))^2$, lo que nos permitirá estimar el modelo con heteroscedasticidad mediante mínimos cuadrados ponderados usando como variable de ponderación $1/log(YD)$. Este ajuste por mínimos cuadrados ponderados es equivalente al ajuste por MCO siguiente:

$$\frac{Log(Q1)}{Log(YD)} = \frac{Log(b_1)}{Log(YD)} + \frac{b_2 Log(p1)}{Log(YD)} + \frac{b_3 Log(P2)}{Log(YD)} + \frac{b_4 Log(P3)}{Log(YD)} + \frac{b_5 Log(YD)}{Log(YD)} + \frac{u}{Log(YD)}$$

Los resultados son los siguientes:

Variable dependiente: Log(Q1)

Parámetro	Estimación	Error estándar	T	P-Valor
CONSTANTE	8,65227	2,38956	3,62086	0,0013
Log(P1)	-0,89142	0,21384	-4,16867	0,0003
Log(P2)	0,55254	0,12936	4,27115	0,0002
Log(P3)	-0,34806	0,18153	-1,91738	0,0667
Log(YD)	0,74123	0,13366	5,54532	0,0000

R-cuadrado = 66,7002 porcentaje
R-cuadrado (ajustado para g.l.) = 61,3723 porcentaje
Error estándar de est. = 0,0241685
F = 12,52 (P-valor = 0,0000)
Estadístico de Durbin-Watson = 2,47093

Se observa que el ajuste es bueno y que los parámetros estimados difieren levemente de los obtenidos sin corregir la heteroscedasticidad.

Ejercicio 3-6. Un empresario textil se plantea dos posibles modelos con el fin de estimar la productividad de sus factorías, en función de la antigüedad de la maquinaria que alquila (A_i), la experiencia de los trabajadores contratados (E_i) y la calidad del combustible utilizado (D_{ji}), siendo esta última una variable ficticia que toma los valores:

D_{1i} = 1 si el combustible es de calidad baja, cero en caso contrario.

D_{2i} = 1 si el combustible es de calidad media, cero en caso contrario.

D_{3i} = 1 si el combustible es de calidad alta, cero en caso contrario.

I = 1,..., 20. (Factorías).

Los modelos son:

(i) $P_i = \beta_0 + \beta_1 A_i + \beta_2 E_i + \beta_3 D_{1i} + \beta_4 D_{2i} + \beta_5 D_{3i} + u_i$

(ii) $P_i = \beta_1 A_i + \beta_2 E_i + \beta_3 D_{1i} + \beta_4 D_{2i} + \beta_5 D_{3i} + u_i$

a.- Razone qué modelo sería más apropiado, si la estimación se lleva a cabo por MCO.
b.- Interprete los coeficientes de las variables en el modelo (ii).
c.- ¿Se podría plantear un modelo alternativo?

En el modelo (*i*) aparece el problema de multicolinealidad exacta entre el término constante y las tres variables ficticias ya que $D_{1i} + D_{2i} + D_{3i} = 1$, para todo *i*. Por tanto no podríamos aplicar MCO para estimar el modelo.

Una posible vía para evitar la multicolinealidad perfecta en el modelo (*i*) es suprimir el término constante. Dado que $D_{1i} + D_{2i} + D_{3i} = 1$, para todo i, sustituyendo en (i):

$$P_i = \beta_0(D_{1i} + D_{2i} + D_{3i}) + \beta_1 A_i + \beta_2 E_i + \beta_3 D_{1i} + \beta_4 D_{2i} + \beta_5 D_{3i} + u_i =$$
$$= \beta_1 A_i + \beta_2 E_i + (\beta_0 + \beta_3) D_{1i} + (\beta_0 + \beta_4) D_{2i} + (\beta_0 + \beta_5) D_{3i} + u_i$$

$$I = 1,\ldots, 20$$

Si renombramos los coeficientes correspondientes a las variables ficticias obtendríamos la formulación correspondiente al modelo (*ii*), en el que ya no existe problema de multicolinealidad exacta, siendo posible su estimación por MCO.

En el modelo (*ii*) los coeficientes β_i con i = 3, 4, 5, correspondientes a las variables ficticias (que podrían estimarse por MCO) se podrían interpretar como el valor de la productividad de la factoría en la que se utilice combustible de la calidad correspondiente (baja, media o alta) que no esté relacionada linealmente con A_i y E_i. En este caso no distinguimos entre un elemento común y uno diferenciador como en el modelo (*i*). En ambos los coeficientes de las variables A_i, E_i, recogen la variación de la productividad media ante la variación en una unidad de la correspondiente variable independiente.

Para solucionar el problema de multicolinealidad perfecta que aparece en el modelo (*i*), un camino alternativo al modelo (*ii*) es eliminar una de las variables ficticias. Si eliminamos la variable D_{2i} tendremos:

$$\text{(i)} \quad \begin{aligned} P_i &= \beta_0 + \beta_1 A_i + \beta_2 E_i + \beta_3 D_{1i} + \beta_4 (1 - D_{1i} - D_{3i}) + \beta_5 D_{3i} + u_i = \\ &= (\beta_0 + \beta_4) + \beta_1 A_i + \beta_2 E_i + (\beta_3 + \beta_4) D_{1i} + (\beta_5 + \beta_4) D_{3i} + u_i \end{aligned}$$

$$I = 1,\ldots, 20$$

Renombrando los coeficientes:

$$P_i = \beta^*_0 + \beta_1 A_i + \beta_2 E_i + \beta^*_3 D_{1i} + \beta^*_4 D_{3i} + u_i$$
$$I = 1,\ldots, 20.$$

que podemos estimar por MCO.

En este caso la interpretación de los coeficientes sería:

$\beta^*_0 = \beta_0 + \beta_4$ representa la productividad de las factorías que utilizan combustible de calidad media que no esté linealmente relacionada con A_i y E_i.

$\beta^*_3 = \beta_3 - \beta_4$ recoge el efecto diferenciador en la productividad media de las factorías con combustible de calidad baja respecto de las que utilizan combustible de calidad media, que no esté linealmente relacionada con A_i y E_i.

Para $\beta^*_4 = \beta_5 - \beta_4$, la interpretación sería semejante a la de β^*_3, esto es, recoge el efecto diferenciador en la productividad media de las factorías con combustible de calidad alta respecto de las que utilizan combustible de calidad media que no esté linealmente relacionada con A_i y E_i.

Ejercicio 3-7. Un empresario desea estimar el coste final de elaboración del producto (P_t), en función del precio de la materia prima utilizada (M_t), y de la calidad con que se elabora el mismo ($A1_t$). Para evaluar dicha calidad se basa en los datos mensuales aportados por una empresa consultora, y es de esperar, por tanto, que la calidad esté medida con algún tipo de error. El empresario posee una medición alternativa de la calidad del producto, proporcionada por otra empresa distinta ($A2_t$). Se utilizan datos mensuales de los últimos tres años, que se presentan a continuación:

Observaciones	P_t	$A1_t$	$A2_t$	M_t	*Observaciones*	P_t	$A1_t$	$A2_t$	M_t
1	25	20	18	32	19	25	20	20	33
2	24	20	19	32	20	24	20	21	33
3	23	20	20	31	21	24	19	20	35
4	24	20	20	32	22	26	20	22	36
5	20	18	19	32	23	24	21	23	30
6	24	19	22	35	24	24	20	23	34
7	25	20	25	34	25	25	20	24	35
8	23	20	21	30	26	22	19	19	32
9	24	19	24	35	27	24	20	20	31
10	20	18	20	32	28	23	19	21	33
11	22	18	19	34	29	25	20	22	34
12	22	19	20	34	30	21	19	22	30
13	25	20	21	36	31	20	19	21	30
14	23	19	22	35	32	25	21	20	31
15	26	20	23	36	33	25	20	20	34
16	24	21	19	31	34	22	19	22	32
17	24	21	19	30	35	23	20	24	32
18	21	19	21	31	36	21	18	23	34

a.- Utilizando un modelo lineal, con variables explicativas $A1_t$, M_t y una constante, estime el coste final del producto mediante la utilización de variables instrumentales, justificando su utilización.
b.- Calcule la matriz de varianzas y covarianzas de los parámetros estimados.
c.- Contraste la hipótesis nula de que el coeficiente de M_t vale 0,3.
d.- Contraste la hipótesis nula de que la suma de los parámetros de las variables M_t y $A1_t$ es igual a la unidad.

a.- Si el coste final del producto depende de la calidad y del precio de la materia prima utilizada, un modelo lineal para estimar dicho coste sería:

$$\text{(i)}\ P_t = \beta_0 + \beta_1 A_t + \beta_2 M_t + u_t \qquad t = 1,\ldots,36$$

donde A_t es el nivel de calidad real del producto. Si utilizamos en la estimación la variable $A1_t$, al medirse esta variable con error, de modo que: $A1_t = A_t + a_t$, tendríamos:

$$P_t = \beta_0 + \beta_1 A1_t + \beta_2 M_t + (u_t - \beta_1 a_t) \qquad t = 1,\ldots,36$$

Y por tanto $A1_t = A_t + a_t$, estará correlacionada contemporáneamente con $v_t = (u_t - \beta_1 a_t)$; si utilizásemos MCO, el estimador del parámetro de $A1$ sería inconsistente. Por ello resulta necesario utilizar variables instrumentales para la estimación. Una posible elección en este caso sería la variable $A2_t$, ya que los errores de medida de $A1$, en principio, no estarán correlacionados con los de $A2$.

Tendríamos así que el estimador por variables instrumentales es:

$$\hat{\beta}_{vi} = (Z'X)^{-1} Z'P \ \text{ dónde } Z_t = [1 \quad A2_t \quad M_t] \ \ X_t = [1 \quad A1_t \quad M_t]$$

siendo los coeficientes estimados:

$$\hat{\beta}_{vi} = \begin{bmatrix} 5{,}004222 \\ 0{,}249519 \\ 0{,}411462 \end{bmatrix}$$

b.- Para hallar la matriz de varianzas y covarianzas debemos buscar primero un estimador consistente de la varianza:

$$\hat{\sigma}_u^2 = \frac{(P - X'\hat{\beta}_{vi})(P - X'\hat{\beta}_{vi})}{T - k - 1} = 1{,}846772$$

(para computar los residuos se deben utilizar las variables originales del modelo, no las instrumentales).

Siendo una estimación consistente de la matriz de varianzas y covarianzas:

$$\hat{V}(\hat{\beta}_{vi}) = \hat{\sigma}_v^2 (Z'X)^{-1}(Z'Z)(X'Z)^{-1}$$

y por tanto tendremos:

$$\hat{V}(\hat{\beta}_{vi}) = 1{,}846772 \begin{bmatrix} 1708{,}388 & -74{,}8977 & -736504 \\ -74{,}8977 & 3{,}300199 & 0{,}313019 \\ -7{,}36504 & 0{,}313019 & 0{,}037649 \end{bmatrix}$$

$$\hat{V}\left(\hat{\beta}_{vi}\right)=\begin{bmatrix} 3155{,}003 & -138{,}319 & -13{,}6016 \\ -138{,}319 & 6{,}094715 & 0{,}578075 \\ -13{,}6016 & 0{,}578075 & 0{,}069529 \end{bmatrix}$$

c.- Para llevar a cabo el contraste de la $H_0 : \beta_2 = 0{,}3$, aplicamos el estadístico:

$$\frac{\hat{\beta}_2 - 0{,}3}{\sqrt{\left(\hat{V}_{\beta_2}\right)}} \approx t_{t-k}$$

El valor del estadístico es 0,422711, menor que el valor crítico de la t correspondiente para 33 grados de libertad igual a 2,7, por lo que no se rechaza la H_0.

d.- Para contrastar la $H_0 : \beta_1 + \beta_2 = 1$, utilizamos el estadístico F, que en este caso se expresa:

$$F = \frac{\left(R\hat{\beta} - r\right)\left[R(Z'X)^{-1}(Z'Z)(X'Z)^{-1}R'\right]^{-1}\left(R\hat{\beta} - r\right)/q}{\hat{\sigma}_u^2} \approx F_{q,t-k}$$

donde $R = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$; r = 1

En nuestro caso tenemos que:

$$F = \frac{0{,}015701}{1{,}846772} = 0{,}015701$$

El valor de las tablas correspondientes a una $F_{1,33}$ al 95% es aproximadamente 4,1, siendo 7,5 al 99%. Por tanto no se rechazaría la hipótesis nula.

CAPÍTULO 4

HERRAMIENTAS PARA TRATAR AUTOCORRELACIÓN, HETEROSCEDASTICIDAD Y OTROS PROBLEMAS

TRATAMIENTO DE LA AUTOCORRELACIÓN Y HETEROSCEDASTICIDAD CON EVIEWS

Eviews permite utilizar prácticamente todas las herramientas para la detección y tratamiento de los problemas de autocorrelación y multicolinealidad expuestos en el capítulo anterior.

Como ejemplo partimos del archivo *economia.wf*1 que contiene los datos de la economía española para el periodo 1985-1997 relativos al consumo público *CP* y al producto interior bruto a precios de mercado *PIBPM*. Con esta información se trata de explicar el consumo público en función del *PIBPM* mediante la ecuación siguiente:

$$CP_t = \beta_0 + \beta_1 PIBPM_t + u_t$$

Se busca ajustar el modelo teniendo presente los efectos de la autocorrelación y que el consumo en un año dado puede depender del consumo en el año anterior.

Si ajustamos el modelo por MCO rellenando la pantalla *Equation Estimation* de Eviews como se indica en la Figura 4-1 se obtienen los resultados de la Figura 4-2.

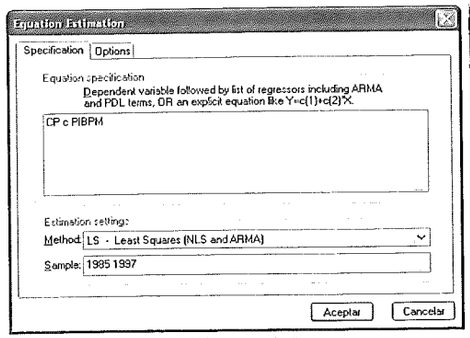


Figura 4-1

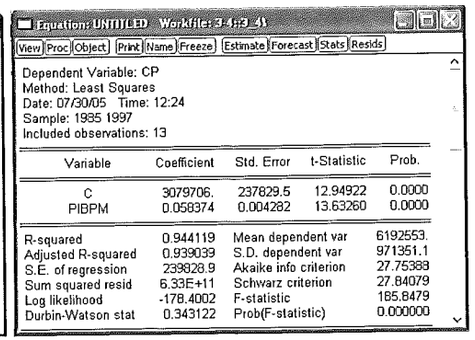


Figura 4-2

Se observa buena significatividad individual y conjunta de los parámetros y buen coeficiente de determinación, pero el valor del estadístico de DurbinWatson inicialmente muy alejado de 2 puede indicar presencia de autocorrelación. En la tabla del estadístico de Durbin-Watson para $T = 13$, $k = 1$ y $\alpha = 0{,}05$ se tiene que $d_L = 1{,}01$ y $d_U = 1{,}34$. Como $d = 0{,}4 < d_L = 1{,}01$ se acepta la existencia de autocorrelación positiva.

Ya sabemos que existe autocorrelación positiva. El siguiente paso será ***detectar el orden de esta autocorrelación***. Para ello puede utilizarse el correlograma residual. Desde la pantalla *Equation*, utilizamos *View* → *Residual Test* → *Correlogram-Q-Statistics* (Figura 4-3). Se eligen 11 retardos (valor por defecto) para representar las funciones de autocorrelación (FAC) y autocorrelación parcial (FACP) de los residuos (Figura 4-4) y al pulsar *OK* se obtienen las gráficas de la Figura 4-5. La función de autocorrelación (correlograma) presenta una estructura de decrecimiento sinusoidal y la función de autocorrelación parcial sólo tiene el primer término significativamente distinto de cero (la única barra que atraviesa las bandas de confianza rayadas). Por tanto los residuos presentan una estructura autorregresiva de orden uno AR(1).

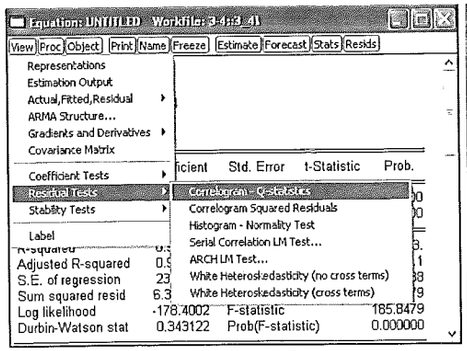


Figura 4-3

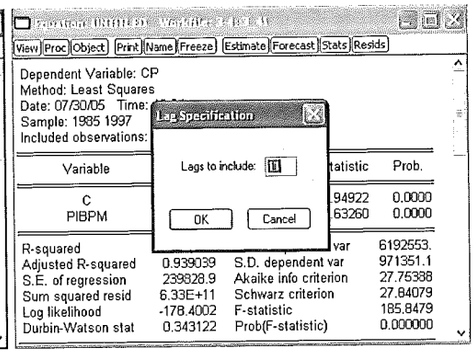


Figura 4-4

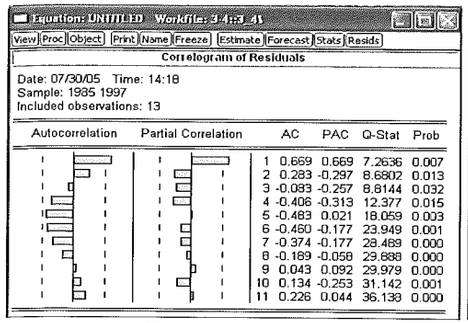


Figura 4-5

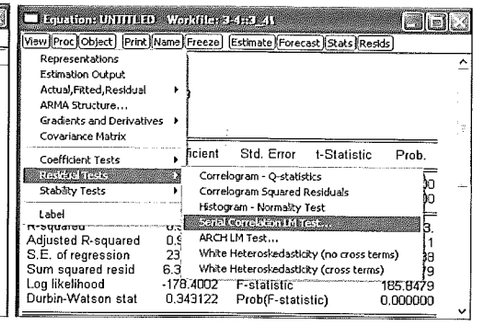


Figura 4-6

También se puede ***detectar la autocorrelación a través del contraste de Breus-Godfrey***. Para ello, desde la pantalla *Equation*, utilizamos *View* → *Residual Tests* → *Serial Correlation LM Test* (Figura 4-37). Se elige 1 retardo en la Figura 4-7 porque ya sabemos que la estructura de los residuos es AR(1) y al pulsar *OK* se obtienen los resultados de la Figura 4-8. Se observan p-valores para la F (0,001498) y la Chi-cuadrado (0,003604) menores que 0,05, lo que nos lleva a aceptar la presencia de autocorrelación. Además, el coeficiente AR(1) estimado es significativo con p-valor = 0,0015 < 0,05, lo que confirma la estructura AR(1) para los residuos.

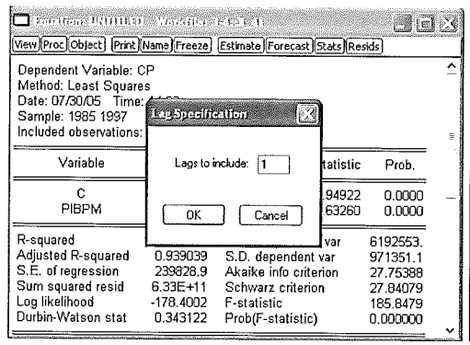


Figura 4-7

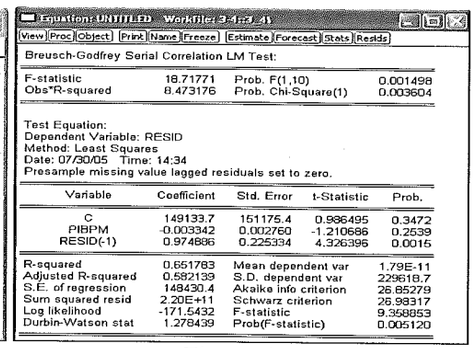


Figura 4-8

En presencia de autocorrelación se puede realizar la ***estimación mediante el método de Cochrane-Orcut con Eviews***. Se elige *Quick* → *Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification* (obsérvese la inclusión del término AR(1) como una variable explicativa más justificada por la presencia de autocorrelación de orden 1), se elige *Least Squares* en el campo *Method* para ajustar por mínimos cuadrados (Figura 4-9) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-10.

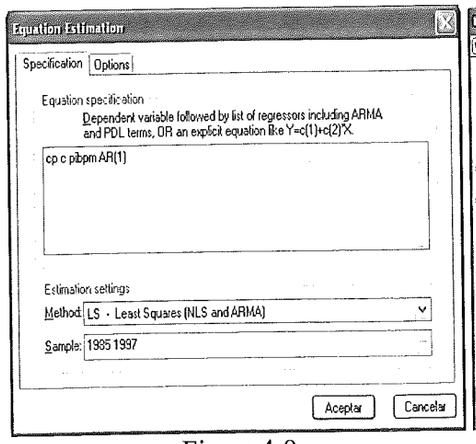


Figura 4-9

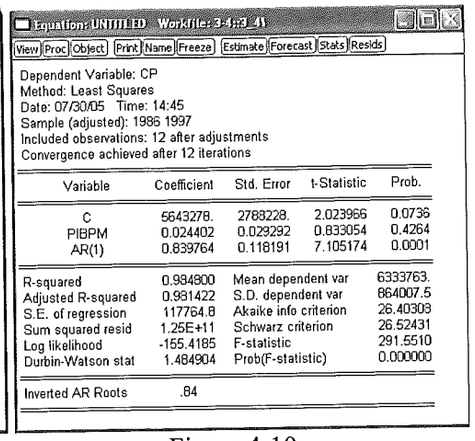


Figura 4-10

Se observa que el parámetro relativo a la variable *PIBPM* no es significativo, con lo que hemos de buscar otra vía para arreglar el problema de la autocorrelación en el modelo.

Como el consumo en un año dado puede depender del consumo en el año anterior, si introducimos en el modelo un retardo de la variable dependiente *CP*(-1), tendríamos para analizar el modelo siguiente:

$$CP_t = \beta_0 + \beta_1 PIBPM_t + \beta_2\, CP_{t-1} + u_t$$

Si ajustamos el nuevo modelo por MCO rellenando la pantalla *Equation Estimation* de Eviews como se indica en la Figura 4-11 se obtienen los resultados de la Figura 4-12. Se observa buen R^2, buena significatividad conjunta y mala significatividad de *PIBPM*.

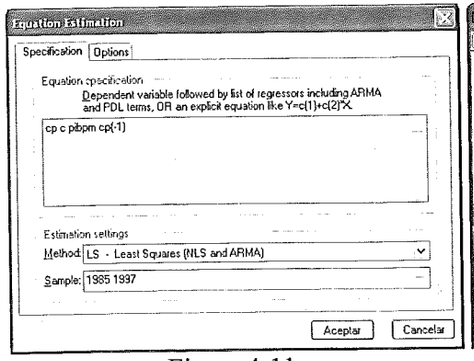


Figura 4-11

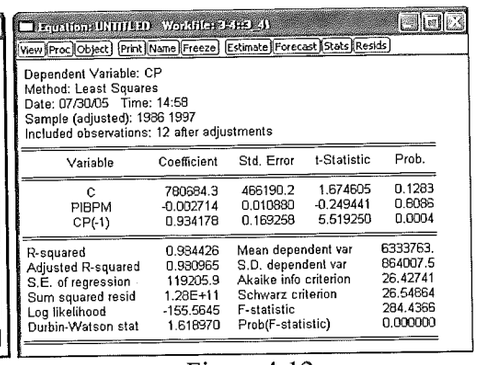


Figura 4-12

Se obtiene un valor del estadístico de Durbin Watson $d = 1{,}61$ muy próximo a 2, pero como el modelo tiene como regresor el regresando desfasado, no es adecuado utilizar el estadístico de Durbin-Watson para analizar la autocorrelación. En su lugar utilizaremos el ***contraste h de Durbin***, cuyo estadístico se define como:

$$h = \hat{\rho}\sqrt{\frac{T}{1-T(\hat{V}(\hat{\beta}_i))}} = \hat{\rho}\sqrt{\frac{T}{1-T(\hat{\sigma}(\hat{\beta}_i))^2}}$$

Tenemos:

$$d = 2(1-\hat{\rho}) \Rightarrow \hat{\rho} = 1-\frac{d}{2} = 0{,}19 \Rightarrow h = \hat{\rho}\sqrt{\frac{T}{1-T(\hat{\sigma}(\hat{\beta}_i))^2}} = (0{,}19)\sqrt{\frac{30}{1-12(0{,}169258)^2}} = 0{,}8$$

Como $|0{,}8| < 1{,}96$ se acepta la hipótesis nula de no autocorrelación para un nivel de significación de 0,05. Vemos entonces que el modelo con un retardo de la variable dependiente está libre de autocorrelación.

También se puede ***detectar la autocorrelación a través del contraste de Breus-Godfrey***. Para ello, desde la pantalla *Equation*, utilizamos *View* → *Residual Test* → *Serial Correlation LM Test* (Figura 4-13). Se elige 1 retardo para la estructura de los residuos y al pulsar *OK* se obtienen los resultados de la Figura 4-14. Se observan p-valores para la *F* (0,751334) y la Chi-cuadrado (0,689866) mayores que 0,05, lo que nos lleva a aceptar la ausencia de autocorrelación.

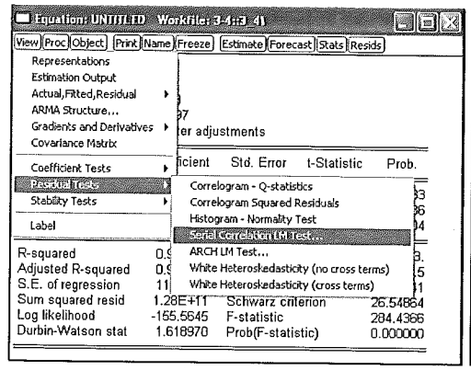


Figura 4-13

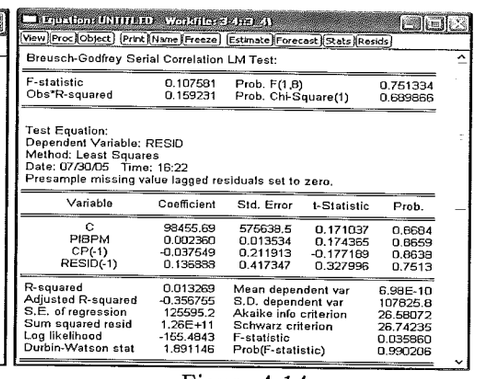


Figura 4-14

Eviews permite realizar el ***contraste de White de heteroscedasticidad*** desde la pantalla *Equation* mediante *View* → *Residual Tests* → *White Heteroskedasticity* (*no cross terms*) (Figura 4-15). Se obtiene la Figura 4-16 en la que se observa que los p-valores de la *F* y de los términos cruzados son mayores que 0,05, lo que nos lleva a aceptar formalmente la ausencia de heteroscedasticidad al 95%.

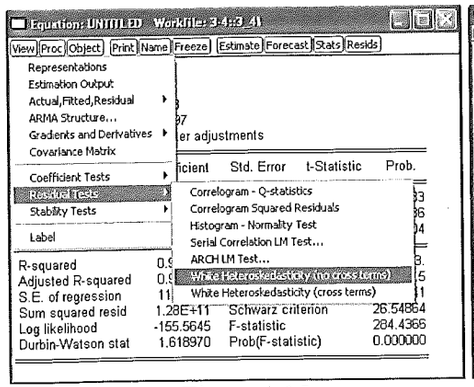


Figura 4-15

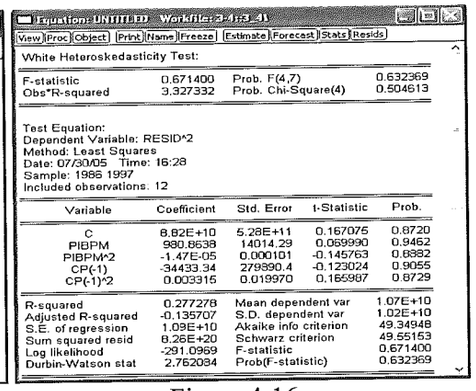


Figura 4-16

Por tanto, el único problema que presenta el modelo con el consumo retardado es la mala significatividad de *PIBPM*. El ajuste sería:

$$CP_t = 780684{,}3 - 0{,}02714\ PIBPM_t + 0{,}934178\ CP_{t-1} + u_t$$

Si ***ajustamos el modelo sin constante por MCO*** rellenando la pantalla *Equation Estimation* de Eviews como se indica en la Figura 4-17 se obtienen los resultados de la Figura 4-18 que presenta un buen ajuste (R^2 alto y significatividad individual y conjunta de los parámetros estimados muy alta).

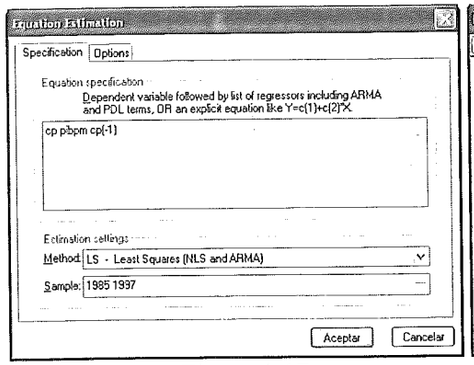


Figura 4-17

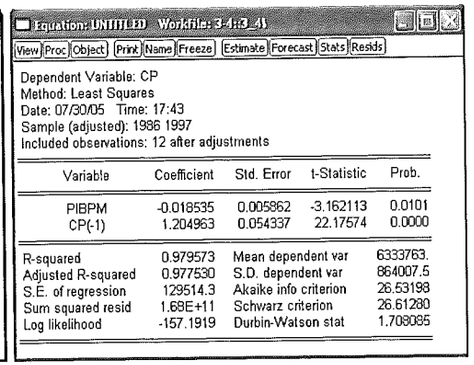


Figura 4-18

Como el modelo no tiene constante, el estadístico de Durbin-Watson no es adecuado para contrastar la autocorrelación, tarea que se lleva a cabo mediante el test de Breusch-Godfrey (Figuras 4-19 y 4-20) cuyos p-valores de la F y la Chi-cuadrado son mayores de 0,005, lo que indica ausencia de autocorrelación. Los p-valores de la F y la Chi-cuadrado mayores que 0,05 del test de White (Figura 4-21) indican ausencia de heteroscedasticidad y el p-valor del test de Jarque-Vera mayor que 0,005 (Figura 4-22) indica normalidad de los residuos.

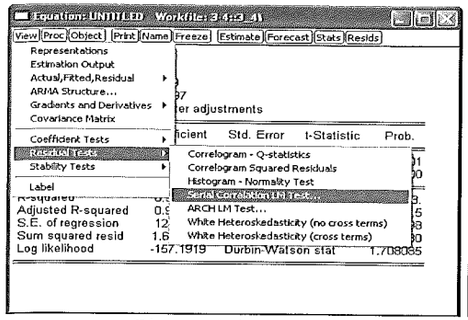


Figura 4-19

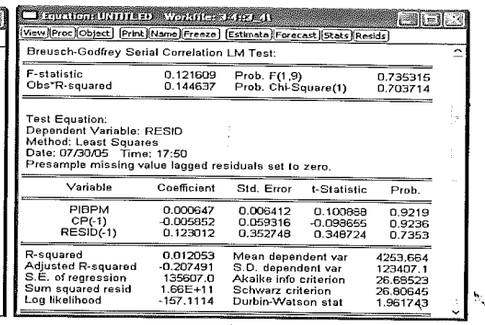


Figura 4-20

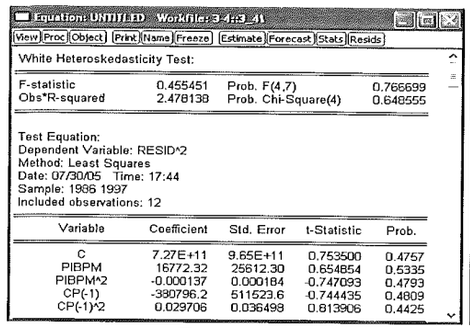


Figura 4-21

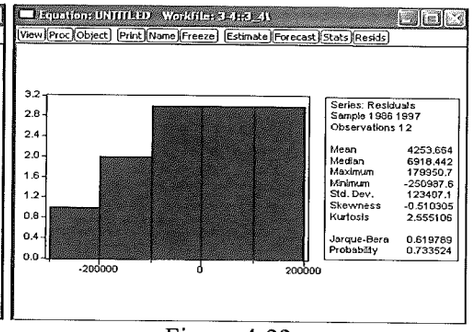


Figura 4-22

El ajuste del modelo con el consumo retardado sin constante será:

$$CP_t = -\,0{,}018535\ PIBPM_t + 1{,}204963\ CP_{t-1} + u_t$$

EVIEWS Y LOS MODELOS ARCH Y GARCH

Partimos del archivo *ofertam.wf*1, contiene los datos relativos a la oferta monetaria *M*1, producto interior bruto *PIB* y tipos de interés *R* para la economía canadiense con datos trimestrales desde enero de 1978 a diciembre de 1987. Con esta información se trata de explicar la demanda de dinero mediante la ecuación siguiente:

$$Ln\ M1 = \beta_0 + \beta_1 LnPIB + \beta_2 LnR + u$$

ajustando el modelo teniendo presente los efectos de la posible heteroscedasticidad condicional en caso de existir.

Si ajustamos el modelo por MCO rellenando la pantalla *Equation Estimation* de Eviews como se indica en la Figura 4-23 se obtienen los resultados de la Figura 4-24.

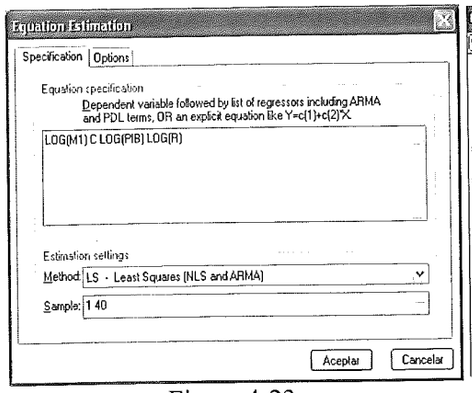


Figura 4-23

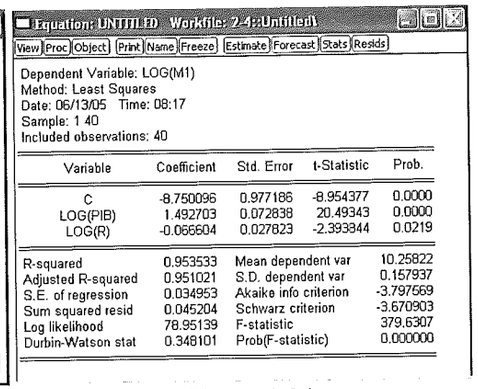


Figura 4-24

Inicialmente el ajuste presenta significatividad al menos al 97% (p-valores de la T menores que 0,03) de todos los parámetros y también hay fuerte significatividad global (p-valor de la F casi nulo). Por otra parte puede existir autocorrelación, ya que el estadístico de Durbin-Watson se aleja mucho de 2. El R^2 es muy alto (superior al 95%).

A continuación ***contrastamos la normalidad de los residuos*** utilizando *View → Residual Tests → Histogram – Normality Test* desde la pantalla *Equation* (Figura 4-25). Se obtienen los resultados de la Figura 4-26.

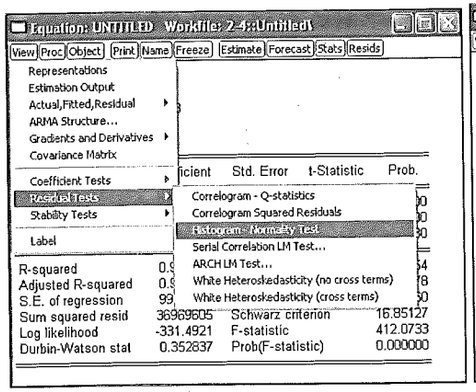


Figura 4-25

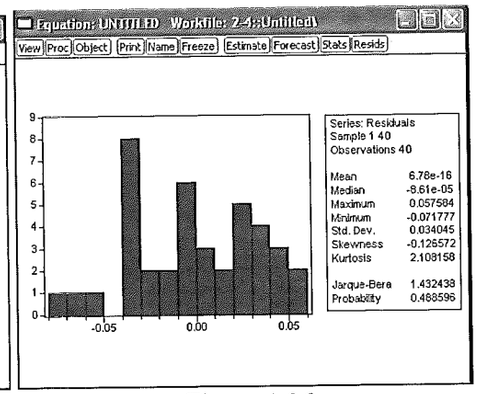


Figura 4-26

En la Figura 4-26 se ve que los coeficientes de asimetría y curtosis residual están prácticamente en la frontera del intervalo [-2 y 2] para poder aceptar la normalidad con reparos, circunstancia que también muestra el histograma. El contraste formal de Jarque-Bera acepta formalmente la normalidad de los residuos (p-valor mayor que 0,05).

Para ***contrastar la posible heteroscedasticidad condicional***, realizamos previamente el correlograma de los residuos al cuadrado utilizando *View → Residual Tests → Correlogran Squared Residuals* desde la pantalla *Equation* (Figura 4-27). Se obtiene el correlograma de la Figura 4-28, en el que se observan coeficientes significativos en el primer retardo (p-valor de 0,007 claramente menor que 0,05) y posiblemente en el segundo con p-valor de 0,014 también menor que 0,05. El tercer retardo ya no se considera porque su p-valor está cerca de 0,05.

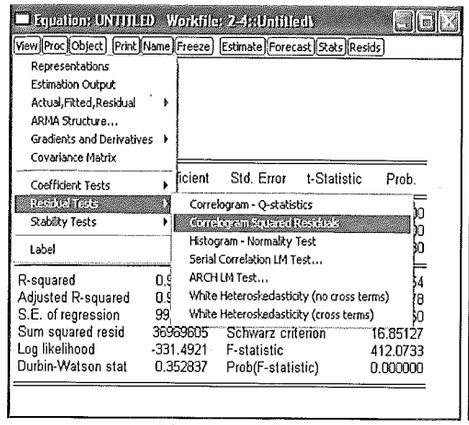


Figura 4-27

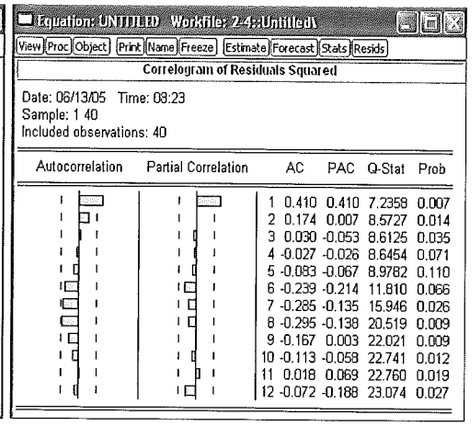


Figura 4-28

Ante esta situación realizamos el contraste formal del multiplicador de Lagrange utilizando dos retardos en la estructura ARCH (p = 2). Para ello se utiliza *View → Residual Tests → ARCH LM Test* desde la pantalla *Equation* (Figura 4-29) con 2 retardos. Se obtienen los resultados de la Figura 4-30 donde, tanto el contraste de la F como el test de Engel del multiplicador de Lagrange ($N*R^2$) aceptan la significatividad de los términos ARCH (p-valores menores que 0,05), lo que muestra la existencia de una cierta heteroscedasticidad condicional.

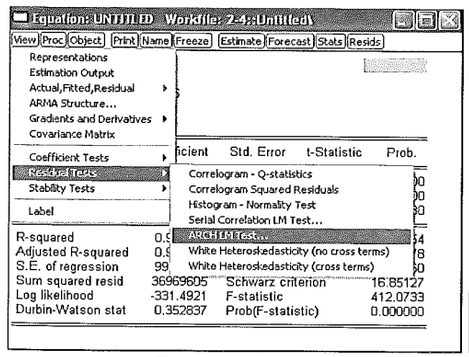


Figura 4-29

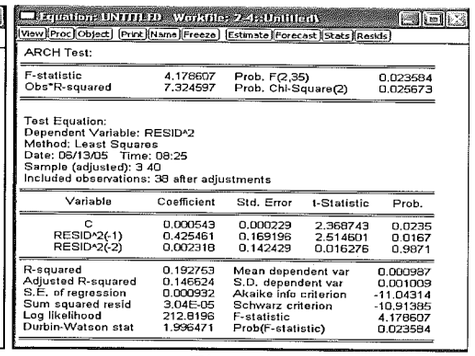


Figura 4-30

La siguiente tarea es realizar la estimación del modelo GARCH(p,q) más adecuado para nuestros datos con *p* y *q* con valores máximos de 2 (derivados de la estructura del correlograma residual de la Figura 4-28). Comenzaremos ajustando un modelo GARCH(2,2) rellenando la solapa *Especification* de la ventana *Equation Estimation* con la ecuación el modelo y eligiendo *ARCH-Autoregressive Conditional Heteroskedasticity* (Figura 4-31). Se obtiene la Figura 4-32 en la que se elige GARCH/TARCH en su campo *Model* introduciendo los órdenes ARCH y GARCH deseados. La solapa *Options* se rellena según se indica en la Figura 4-33. Al pulsar *Aceptar* se obtiene el ajuste al modelo GARCH(2,2) de la Figura 4-34.

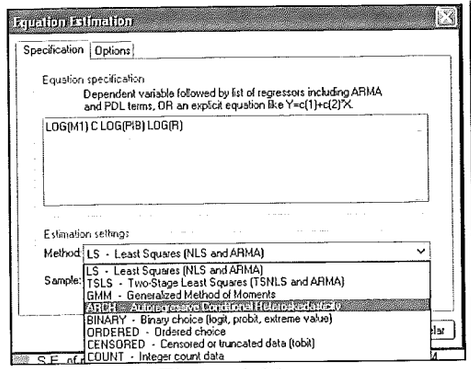


Figura 4-31

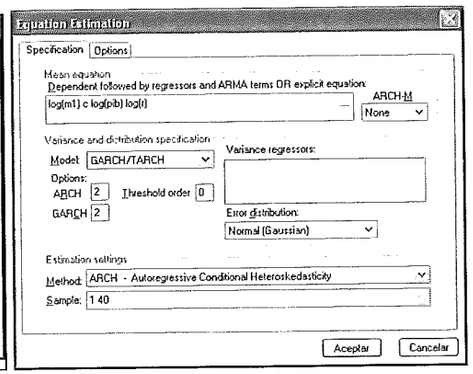


Figura 4-32

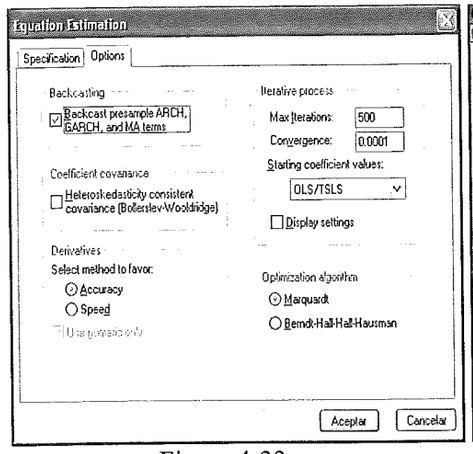


Figura 4-33

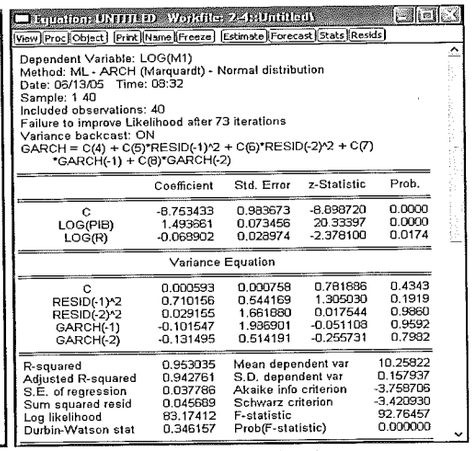


Figura 4-34

Se observa que los coeficientes del GARCH(1) y el GARCH(2) son negativos (en *Variance Equation*), luego habrá que prescindir de estos términos y ajustar un modelo ARCH(2,0) como estructura más completa posible.

Para ello se rellena la solapa *Specification* de la pantalla *Equation Estimation* como se indica en la Figura 4-35. Al pulsar *Aceptar* se obtiene el ajuste de la Figura 4-36 que presenta baja significatividad del coeficiente ARCH(2). Ello nos lleva a considerar la estructura ARCH(1) rellenando la solapa *Specification* de la pantalla *Equation Estimation* como se indica en la Figura 4-36. Al pulsar *Aceptar* se obtiene el ajuste de la Figura 4-37 que presenta significatividades muy altas para los parámetros de las variables del modelo tanto individual como conjunta. La significatividad del término ARCH es casi el 90%.

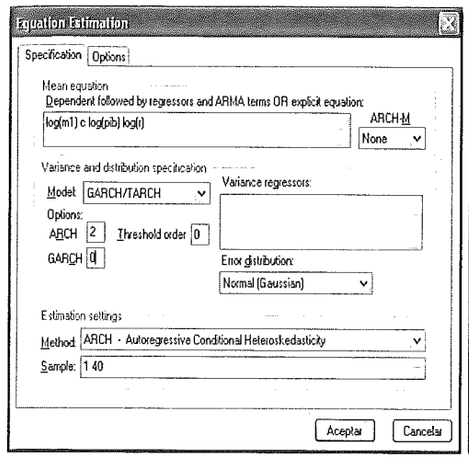


Figura 4-35

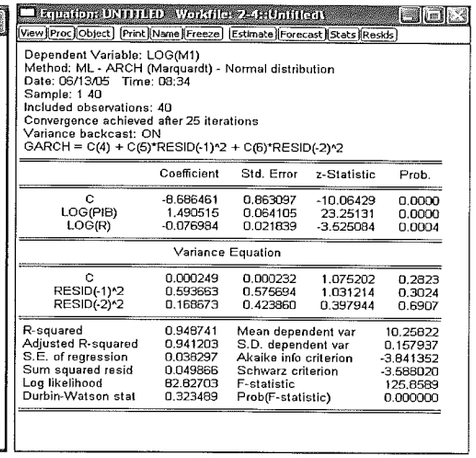


Figura 4-36

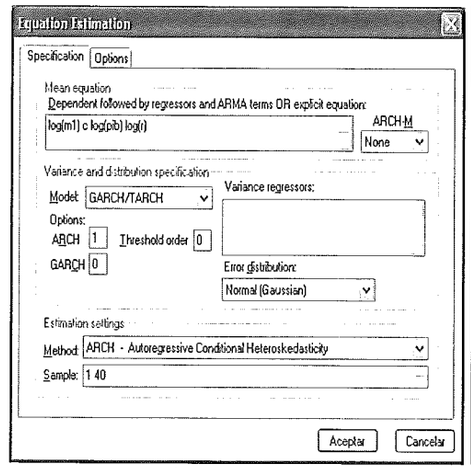


Figura 4-37

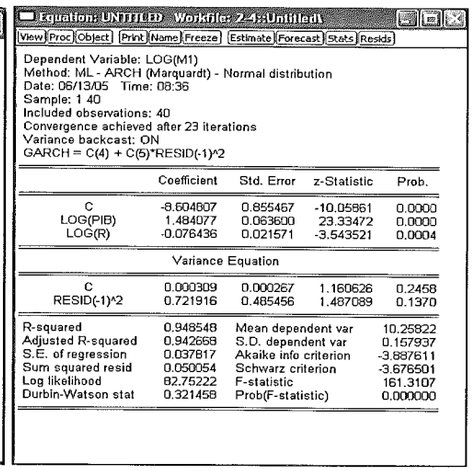


Figura 4-38

También puede intentarse un ***ajuste a un modelo EGARCH (1,0)*** rellenando el campo *Model* como se indica en la Figura 4-39. En la Figura 4-40 se presenta el ajuste que no resulta demasiado significativo.

También puede intentarse un ajuste ***a un modelo PARCH (1,0)*** rellenando el campo *Model* como se indica en la Figura 4-41. En la Figura 4-42 se presenta el ajuste que no resulta demasiado significativo.

También puede intentarse un ajuste ***a un modelo CGARCH (1,1)*** o ***Component ARCH(1,1)***, rellenando el campo *Model* según la Figura 4-43. En la Figura 4-44 se presenta el ajuste que presenta significatividad dudosa de los términos ARCH, pero más aceptable que en los casos anteriores.

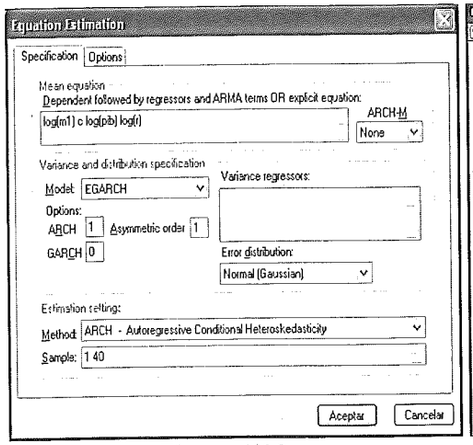


Figura 4-39

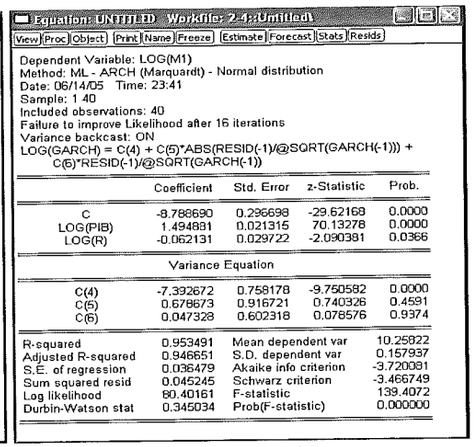


Figura 4-40

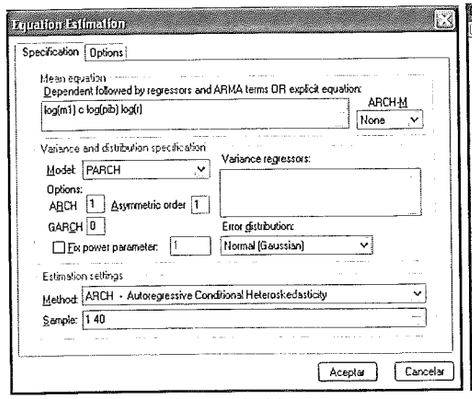


Figura 4-41

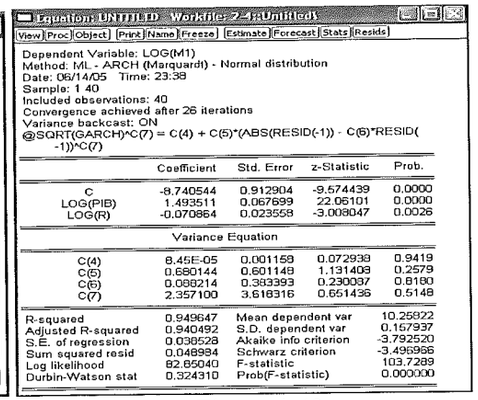


Figura 4-42

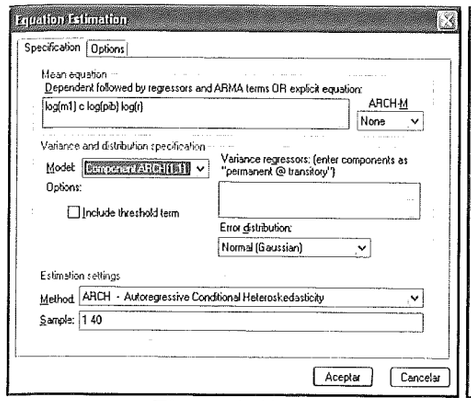


Figura 4-43

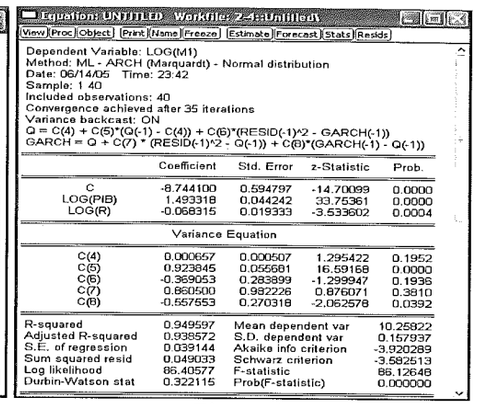


Figura 4-44

Finalmente consideramos que la heteroscedasticidad condicional se ajusta adecuadamente considerando un modelo ARCH(1) o GARCH(1,0). El ajuste final será:

$$Ln\ M1 = -8{,}604807 + 1{,}484077 LnPIB - 0{,}043736 LnR + u$$
$$\sigma_t^2 = V(u_t) = 0{,}000309 + 0{,}721916\ u_{t-1}^{\ 2}$$

ENDOGENEIDAD, VARIABLES INSTRUMENTALES Y MÍNIMOS CUADRADOS EN DOS ETAPAS CON EVIEWS

Partimos del archivo *demanda.wf*1 que contiene datos sobre un modelo de demanda de productos agrícolas. El modelo involucra las variables *DEMANDA, PRECIO, INGRESOS* y *PRECIPIT* que indican el nivel de demanda y el precio del producto, así como los ingresos del consumidor y el nivel de precipitaciones respectivamente.

Se trata de ajustar a los datos un modelo adecuado que explique la demanda del producto en función de su precio y los ingresos del consumidor.

Al tratarse de un modelo de demanda pueden existir problemas de regresores estocásticos, es decir, que los errores de la variable dependiente estén correlacionados con la variable o variables independientes. Para probarlo utilizaremos el contraste de Hausman tomando como instrumento para los precios un retardo de los mismos y para los ingresos el nivel de precipitaciones *precipit*. Está claro que hay correlación entre cada variable y su instrumento (las precipitaciones influyen sobre los ingresos y toda variable está correlacionada con un desfase de la misma).

Para ***probar la endogeneidad*** de los precios aplicamos el ***contraste de Hausman***. Para ello, en un primer paso planteamos la regresión de la variable *precios* respecto a su instrumento *precios*(-1) con el objeto de calcular los residuos *resid01*.

A continuación se elige *Quick → Estimate Equation*. Se rellena la pantalla *Equation Specification* de la solapa *Specification* como se indica en la Figura 4-45, se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar*. Ahora para guardar los residuos en la variable *resid01* elegimos *proc→ Make Residual Series* (Figura 4-46) y escribimos el nombre *resid01* en la Figura 4-47. Al pulsar *OK*, la Figura 4-48 muestra la variable *resid01* como otra variable del conjunto de datos.

A continuación, en un segundo paso, realizamos la regresión de la variable dependiente del modelo inicial *demanda* respecto de la variable *precios* y la variable *resid01*. Se elige *Quick → Estimate Equation*. Se rellena la pantalla *Equation Specification* de la solapa *Specification* como se indica en la Figura 4-49, se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-50 en la que se observa una significatividad muy elevada de la variable *resid01*. Ello nos lleva a aceptar la endogeneidad del regresor *precios* según el contraste de Hausman.

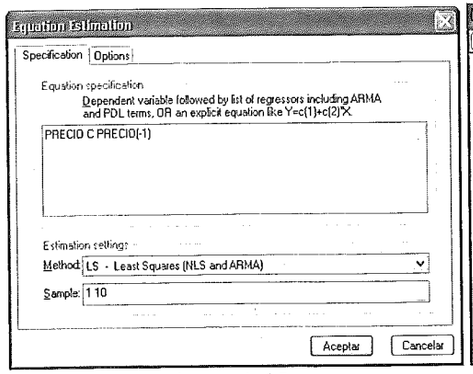


Figura 4-45

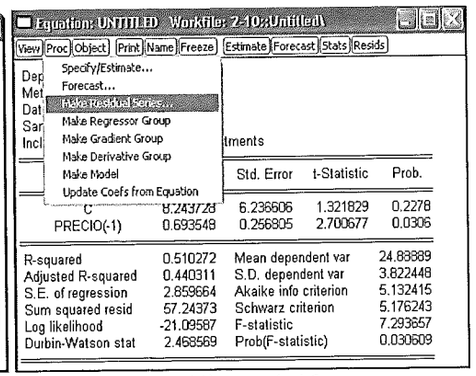


Figura 4-46

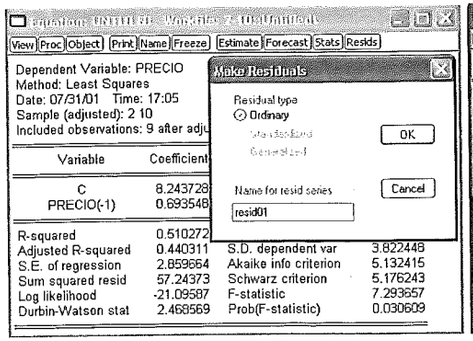


Figura 4-47

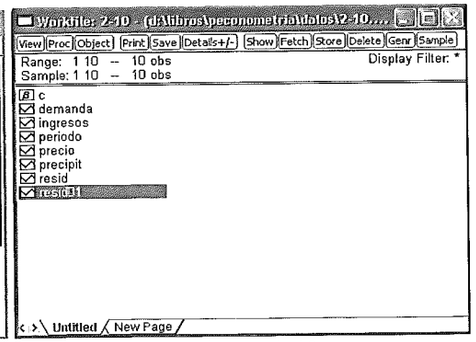


Figura 4-48

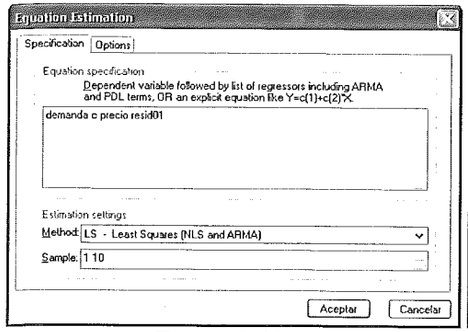


Figura 4-49

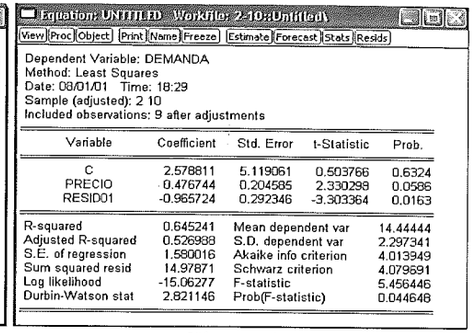


Figura 4-50

Para ***probar la endogeneidad*** de los ingresos mediante el ***contraste de Hausman***, en un primer paso planteamos la regresión de la variable *ingresos* respecto a su instrumento *precipit* con el objeto de calcular los residuos *resid02*. Se elige *Quick → Estimate Equation*. Se rellena la pantalla *Equation Specification* de la solapa *Specification* como se indica en la Figura 4-51, se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar*. Ahora para guardar los residuos en la variable *resid02* elegimos *proc→ Make Residual Series* y escribimos el nombre *resid02* en la Figura 4-52. Al pulsar *OK*, la variable *resid02* es ya otra variable del conjunto de datos.

A continuación, en un segundo paso, realizamos la regresión de la variable dependiente del modelo inicial *demanda* respecto de la variable *ingresos* y la variable *resid02*. Se elige *Quick → Estimate Equation*. Se rellena la pantalla *Equation Specification* de la solapa *Specification* como se indica en la Figura 4-53, se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-54 en la que se observa una significatividad muy elevada de la variable *resid02*. Ello nos lleva a aceptar la endogeneidad del regresor *ingresos* según el contraste de Hausman.

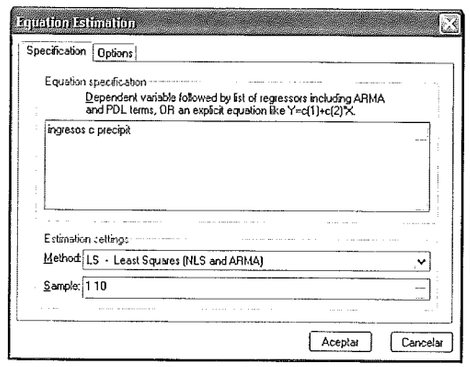


Figura 4-51

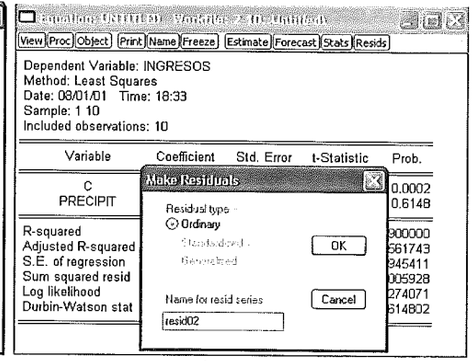


Figura 4-52

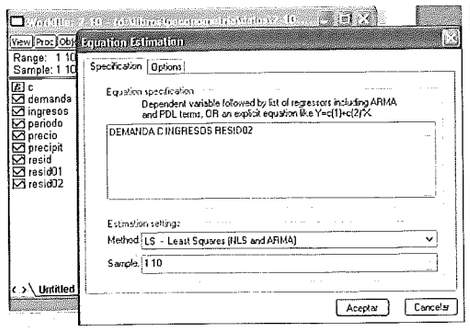


Figura 4-53

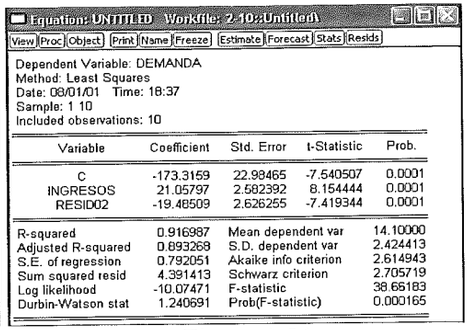


Figura 4-54

Ahora ajustaremos el modelo utilizando ***variables instrumentales*** mediante el ***método de mínimos cuadrados en dos etapas***. Se tendrá en cuenta que hay que introducir como instrumentos por lo menos las variables exógenas y sus instrumentos. Para realizar el ajuste por mínimos cuadrados en dos etapas en Eviews se elige *Quick → Estimate Equation*. Se rellena la pantalla *Equation Specification* de la solapa *Specification* como se indica en la Figura 4-55 seleccionando en el campo *Method* la opción *TSLS – Two-Stage Least Squares (TSNLS and ARMA)*. En el campo *Instruments List* se introducen como instrumentos las variables exógenas y sus instrumentos. En el campo *Equation Specification* se introducen las variables del modelo a ajustar. Al hacer clic en *Aceptar* se obtienen los resultados de la Figura 4-56.

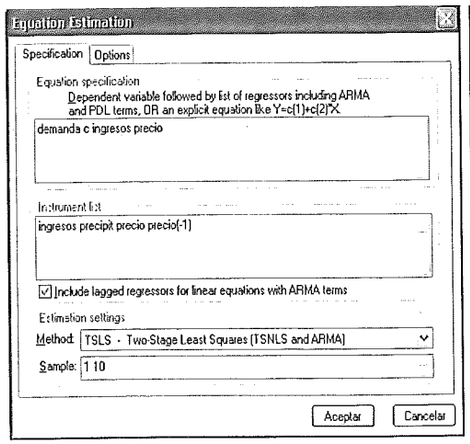


Figura 4-55

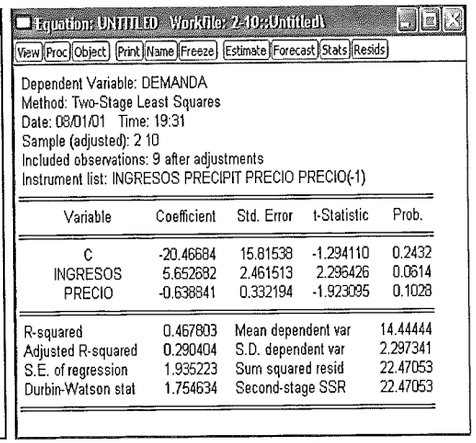


Figura 4-56

La significatividad de las variables es como poco del 90% y la significatividad de la constante se aproxima al 80%. Tampoco hay problemas de autocorrelación porque el estadístico de Durbin-Watson se aproxima a 2. El modelo ajustado tiene la siguiente expresión:

$$Demanda = -20{,}46684 + 5{,}652682\ ingresos - 0{,}638641\ precio$$

Lógicamente la demanda es directamente proporcional a los ingresos (signo positivo del coeficiente estimado) e inversamente proporcional al precio (signo negativo del coeficiente estimado), siendo la influencia de los ingresos sobre la demanda del producto bastante superior a la del precio de dicho producto.

Para analizar la posible heteroscedasticidad utilizaremos el ***contraste de White***. Para ello, desde la pantalla *Equation* mediante *View → Residual Tests → White Heteroskedasticity (no cross terms)* (Figura 4-57) se obtiene la Figura 4-58 en la que se observa que los p-valores de la F y de los términos cruzados son mayores que 0,05, lo que nos lleva a rechazar formalmente la presencia de heteroscedasticidad al 95%.

Podemos probar la ***normalidad de los residuos*** desde la pantalla *Equation* mediante *View → Residual Tests → Histogram – Normality Test* (Figura 4-59). En la Figura 4-60 se ve que los coeficientes de asimetría y curtosis residual están prácticamente en la frontera del intervalo [-2 y 2] para poder aceptar la normalidad con reparos, circunstancia que también muestra el histograma. El contraste formal de Jarque-Bera acepta formalmente la normalidad de los residuos (p-valor mayor que 0,05).

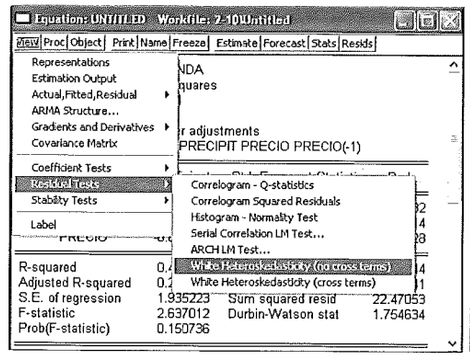


Figura 4-57

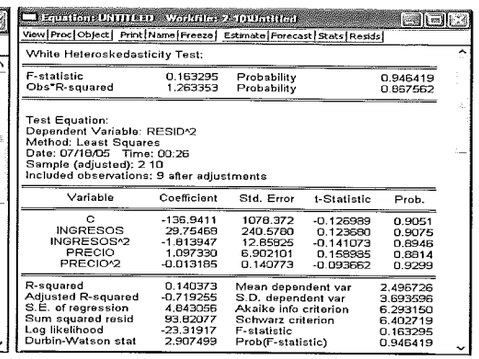


Figura 4-58

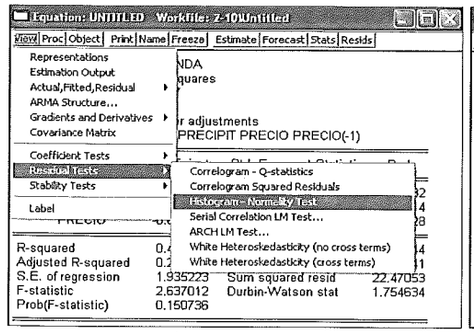


Figura 4-59

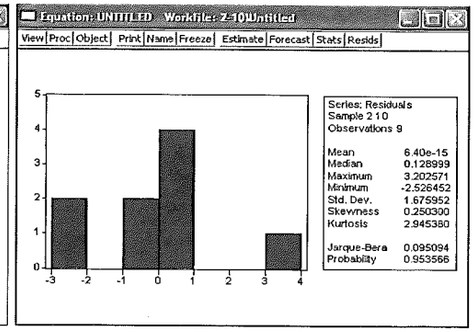


Figura 4-60

ERRORES DE ESPECIFICACIÓN CON EVIEWS. VARIABLES OMITIDAS Y REDUNDANTES

Consideremos una multinacional que trata de analizar la relación existente entre el grado de estrés de los trabajadores Y, medido a partir del tamaño de la sucursal en que trabajan X_1, el número de años que llevan en el puesto de trabajo actual X_2, el salario anual percibido X_3 y la edad del trabajador X_4. Para ello se dispone de las observaciones contenidas en el archivo 2-7.*wf*1. Se trata de obtener el modelo lineal mejor especificado que explique el grado de estrés de los trabajadores en función de las restantes variables y que se ajuste a los datos dados.

Inicialmente consideraremos el modelo completo. Para realizar el ajuste lineal con Eviews, se elige *Quick* → *Estimate Equation*. Se rellena la pantalla *Equation Specification* de la solapa *Specification* como se indica en la Figura 4-61), se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-62.

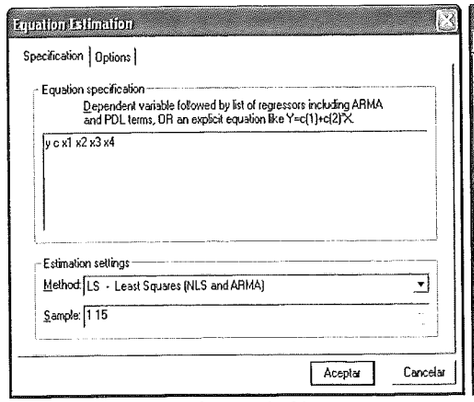


Figura 4-61

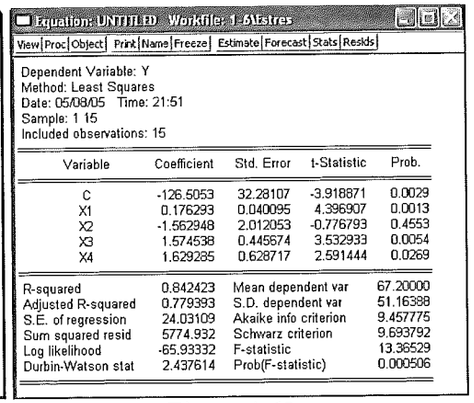


Figura 4-62

Se observa que todos los parámetros estimados (salvo el coeficiente de X_2) son significativamente distintos de cero al 95% (p-valores menores que 0,05) y que la significatividad conjunta es muy alta (p-valor de la F muy pequeño). El R^2 y el R^2 ajustado son 0,84 y 0,78 respectivamente (muy altos) con lo que la variabilidad explicada es alta. La desviación típica estimada del error es 24,03 y los criterios de información de Akaike y Schwartz tienen valores pequeños con lo que la capacidad explicativa del modelo es buena. El estadístico de Durbin Watson tiene un valor no demasiado alejado de 2, lo que indica ausencia de posibles problemas de autocorrelación.

Dado que la variable X_2 ha resultado no significativa individualmente, comprobaremos si es una ***variable redundante a través del ratio de verosimilitud*** y hay que eliminarla de la especificación correcta del modelo. Para ello, sobre la pantalla *Equation*, elegimos *View* → *Coefficient Tests* → *Redundant Variables – Likelihood Ratio* (Figura 4-63) y situamos la variable X_2 en el campo *One or more test series* de la Figura 4-64. Al pulsar *OK* se obtiene la salida de la Figura 4-63. Dado que los p-valores de la F y de la razón de verosimilitud son mayores que 0,05 aceptamos la hipótesis de que la variable X_2 es redundante. Además, la Figura 4-65 presenta el modelo ajustado sin la presencia de la variable X_2.

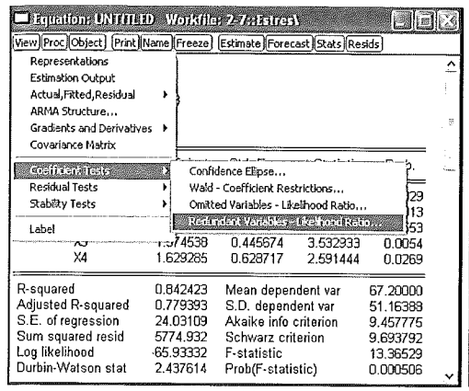


Figura 4-63

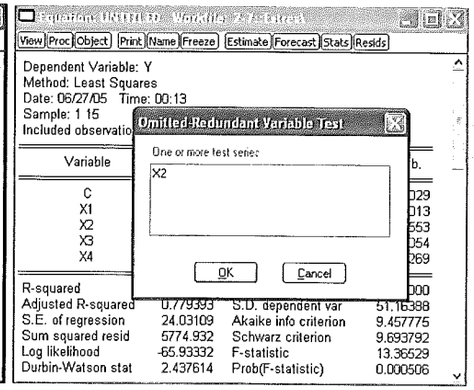


Figura 4-64

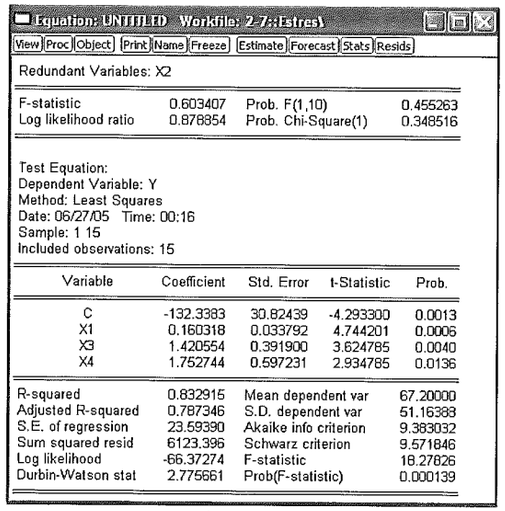


Figura 4-65

En la Figura 4-65 se observa que todas las variables del modelo han aumentado su significatividad (p-valores más pequeños) y los valores de R^2 son mejores (más altos) al eliminar la variable X_2 del modelo. La significatividad conjunta de los parámetros del modelo también ha mejorado (p-valor de la F más pequeño). Observando los coeficientes estimados se deduce que un aumento en una unidad la variable X_1 provoca un aumento en Y de 0,16 unidades suponiendo constantes el resto de las variables. Un aumento de una unidad en la variable X_3 con las demás variables constantes provoca un aumento en Y de 1,42 unidades. Para la variable X_4 el aumento es superior (1,75).

Se ve que al eliminar del modelo la variable X_2, los efectos de las variables independientes sobre la variable dependiente han disminuido levemente. Por lo tanto, la introducción de la variable redundante en el modelo provocaba una sobreestimación de los efectos de las restantes variables sobre la variable dependiente.

No obstante, para comprobar que el modelo sin la variable X_2 está bien especificado definitivamente, será necesario comprobar que X_2 no es una ***variable omitida a través del ratio de verosimilitud***. Para ello, sobre la pantalla *Equation*, elegimos *View* → *Coefficient Tests* → *Omitted Variables – Likelihood Ratio* (Figura 4-66) y situamos la variable X_2 en el campo *One o more test series* de la Figura 4-67. Al pulsar *OK* se obtiene la salida de la Figura 4-68.

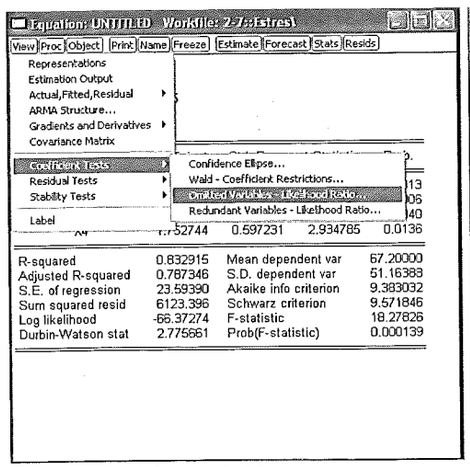


Figura 4-66

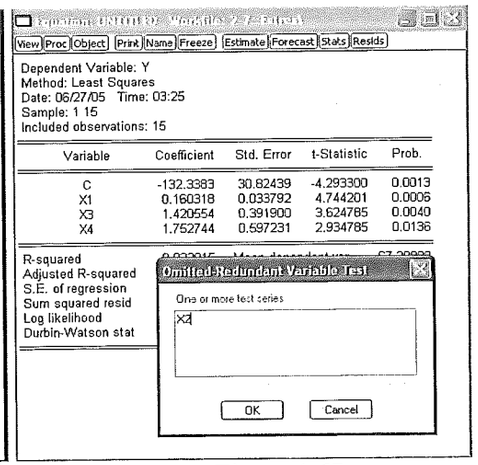


Figura 4-67

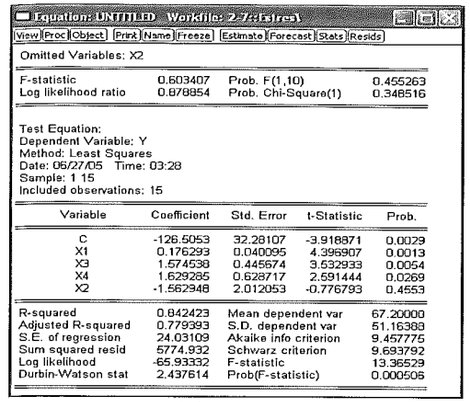


Figura 4-68

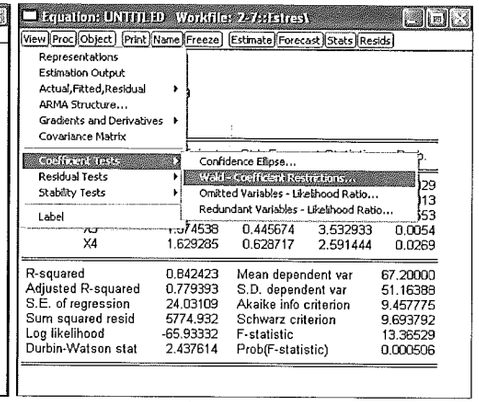


Figura 4-69

Las altas probabilidades de la F y del logaritmo del ratio de verosimilitud indican que hay que aceptar la hipótesis de que la variable X_2 no es relevante para explicar la variable endógena. Podemos por tanto omitirla en el modelo.

El ajuste final del modelo será el especificado en la Figura 2-100.

Hubiera sido equivalente el contraste de redundancia (u omisión) de X_2 a haber contrastado la ***significatividad del regresor individual X_2 mediante el test de Wald*** obtenido mediante *View* → *Coefficient Test* → *Wald – Coefficient Restrictions* (Figura 4-69) y situando la nulidad del parámetro C(3) de la variable X_2 en la pantalla *Wald Test* de la Figura 4-70. Al pulsar *OK* se obtiene la salida de la Figura 4-71. Las altas probabilidades de la F y del logaritmo del ratio de verosimilitud indican que hay que aceptar la hipótesis de que la variable X_2 puede considerarse nula para explicar la variable endógena. Podemos por tanto omitirla definitivamente en el modelo.

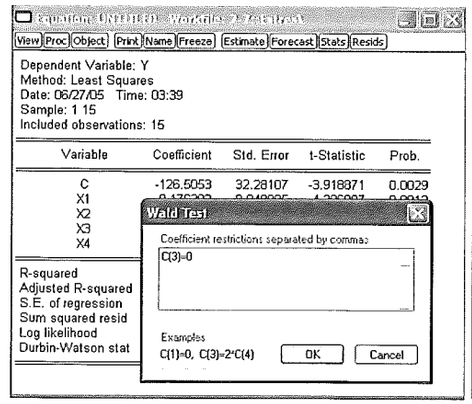


Figura 4-70

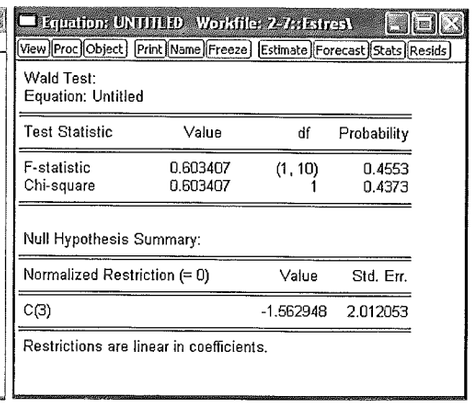


Figura 4-71

ERRORES DE ESPECIFICACIÓN EN LA FORMA FUNCIONAL CON EVIEWS

Consideramos el archivo 2-9.wf1 que contiene datos para la economía española desde 1970 a 1993 sobre los ingresos del sector público INGPUB, el PIB, el ahorro privado AHORRO, las importaciones IMPORT y las exportaciones EXPORT. Se trata de encontrar un modelo adecuado que explique los ingresos del sector público en función del PIB, el ahorro privado, las importaciones y las exportaciones.

Inicialmente se ajusta el modelo por mínimos cuadrados ordinarios usando *File → Open → Eviews Workfile...* y eligiendo *Quick → Estimate Equation*. Se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification* (Figura 4-72), se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-73. Se observa que todos los parámetros estimados (incluida la constante) son significativamente distintos de cero (p-valores muy pequeños que indican una significatividad individual muy alta para los parámetros). La significatividad conjunta también es muy alta porque el p-valor de la F es prácticamente nulo. El R^2 y el R^2 ajustado son 0,991 y 0,99 respectivamente (muy altos).

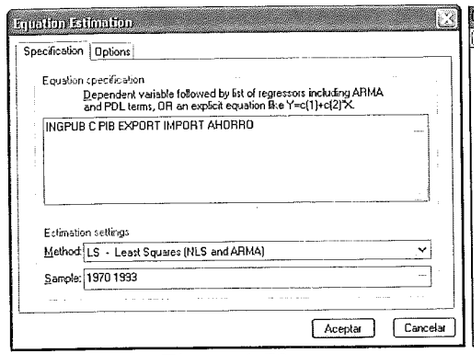


Figura 4-72

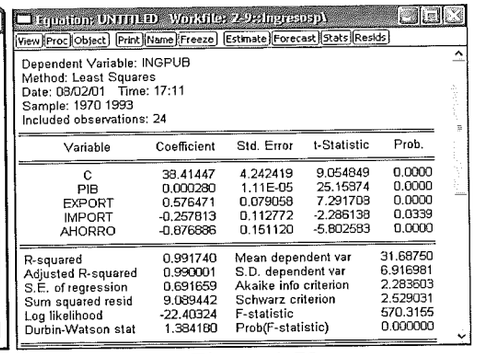


Figura 4-73

Sin embargo, el valor del estadístico de Durbin-Watson está demasiado alejado del valor 2, lo que podría provocar problemas de autocorrelación.

Ante este problema, sería interesante también estudiar una posible mala especificación funcional contrastando los ***posibles problemas de no linealidad***. Para ello realizamos el gráfico de residuos contra las variables explicativas, que han de ser aleatorios. Para ello elegimos *Quick → Graph → Scatter* (Figura 4-74) y rellenamos la pantalla *Series List* como se indica en la Figura 4-75 con el objeto de graficar los residuos contra el PIB.

Al pulsar *OK* se obtiene la Figura 4-76. Se observa que la distribución de los puntos del gráfico no es aleatoria (la nube de puntos podría ser ajustable por una parábola). Ello nos lleva a ensayar ***contrastes formales de especificación del modelo***, por ejemplo el ***test RESET de Ramsey***. Para ello elegimos *View → Stability Tests → Ramsey RESET* (Figura 4-77) y tomamos una sola potencia para la variable endógena ajustada (Figura 4-78). Al pulsar *OK* se obtiene la salida del test RESET de Ramsey de la Figura 4-79. Los p-valores de la F y de la Chi-cuadrado de razón de verosimilitud son menores que 0,05, con lo que rechazamos la hipótesis de linealidad, o más generalmente, de especificación correcta del modelo.

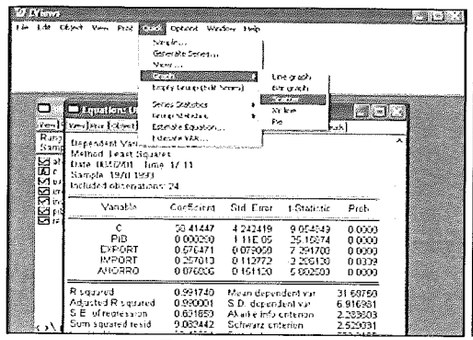


Figura 4-74

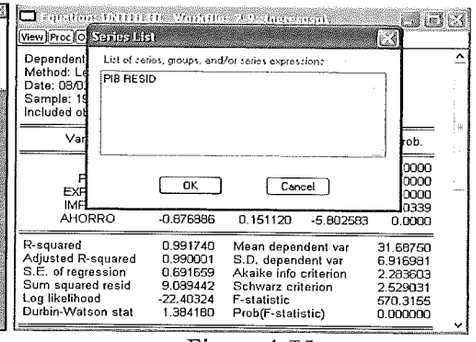


Figura 4-75

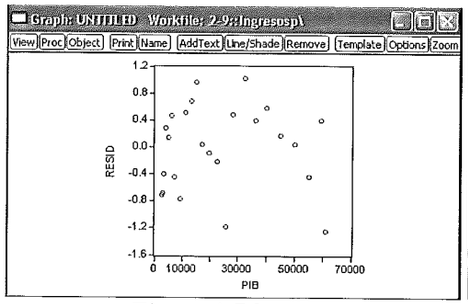


Figura 4-76

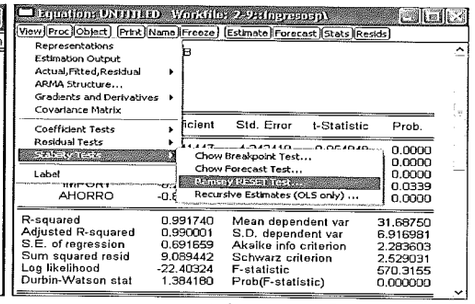


Figura 4-77

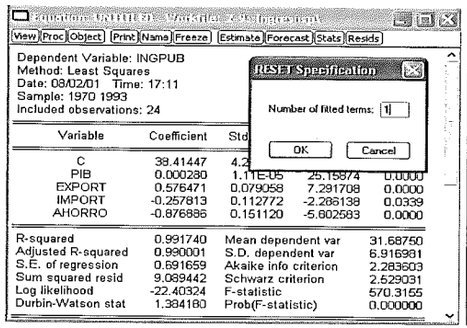


Figura 4-78

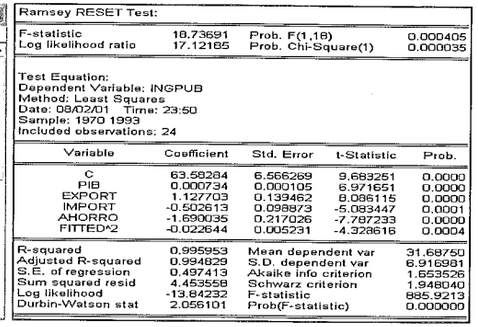


Figura 4-79

Como en la Figura 4-76 se observaba que la nube de puntos de la representación del PIB respecto de los residuos del modelo podía ajustarse por una parábola, ello nos lleva a pensar en introducir el PIB al cuadrado como nueva variable del modelo.

El nuevo modelo se ajusta ahora usando *File → Open → Eviews Workfile...* y eligiendo *Quick → Estimate Equation*. Se escribe la ecuación del modelo en el campo *Equation Specification* de la solapa *Specification* (Figura 4-80), se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-81. Se observa que todos los parámetros estimados (incluida la constante) son significativamente distintos de cero (p-valores muy pequeños que indican una significatividad individual muy alta para los parámetros). La significatividad conjunta también es muy alta porque el p-valor de la F es prácticamente nulo.

El R^2 y el R^2 ajustado son 0,996 y 0,995 respectivamente (muy altos y mejores que en el modelo inicial). Ahora ya no se observan problemas de autocorrelación ya que el estadístico de Durbin-Watson es prácticamente 2.

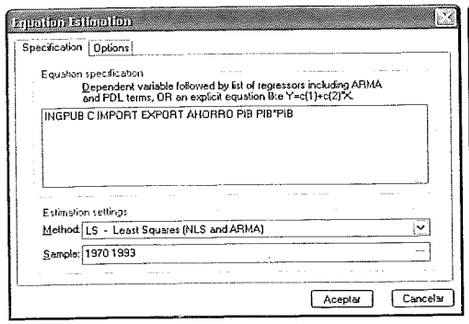


Figura 4-80

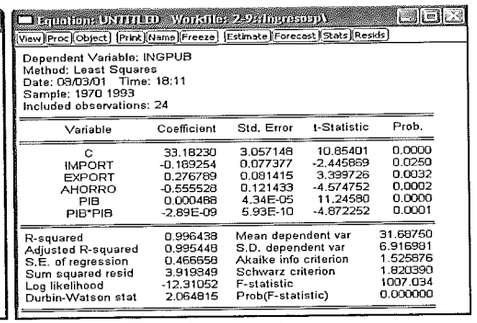


Figura 4-81

Para analizar la posible heteroscedasticidad utilizaremos el ***contraste de White*** consistente en efectuar una regresión de los cuadrados de los residuos MCO sobre todas las variables independientes del modelo, sus cuadrados y sus productos cruzados de dos en dos. La homoscedasticidad se acepta al 95% si los p-valores de la F y de los términos cruzados son mayores que 0,05. Desde la pantalla *Equation* mediante *View → Residual Tests → White Heteroskedasticity* (*no cross terms*) (Figura 4-82) se obtiene la Figura 4-83 en la que se observa que los p-valores de la F y de los términos cruzados son mayores que 0,05, lo que nos lleva a rechazar formalmente la presencia de heteroscedasticidad al 95%.

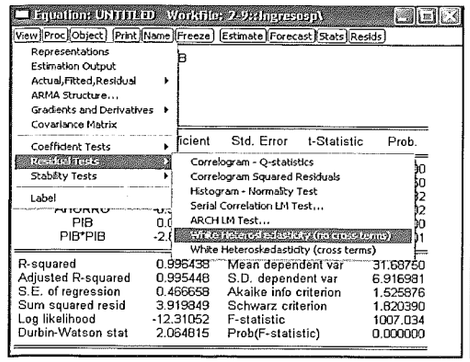


Figura 4-82

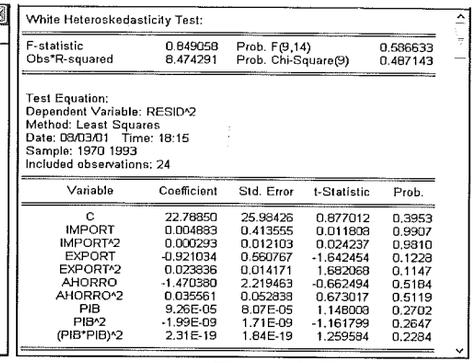


Figura 4-83

Podemos probar la ***normalidad de los residuos*** desde la pantalla *Equation* mediante *View → Residual Test → Histogram – Normality Test* (Figura 4-84). En la Figura 4-85 se ve que los coeficientes de asimetría y curtosis residual están prácticamente en la frontera del intervalo [-2 y 2] para poder aceptar la normalidad con reparos, circunstancia que también muestra el histograma. El contraste formal de Jarque-Bera acepta formalmente la normalidad de los residuos (p-valor mayor que 0,05). Mediante *View → Representation* se obtiene el modelo final ajustado (Figura 4-86).

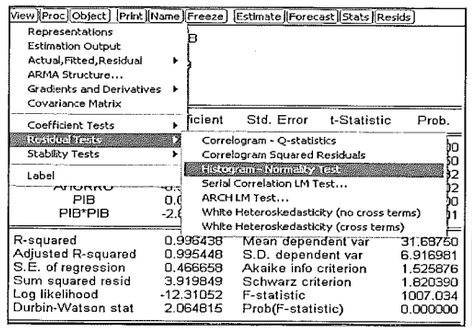


Figura 4-84

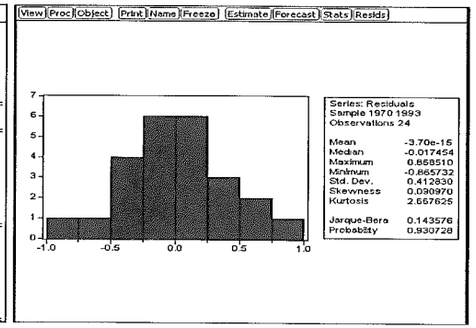


Figura 4-85

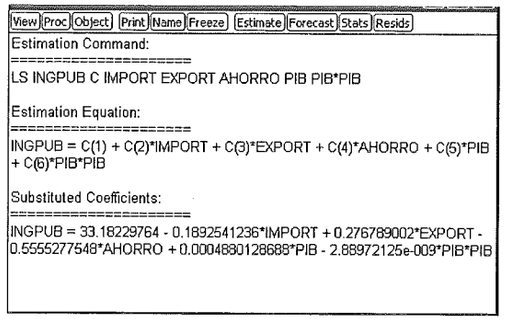


Figura 4-86

SPSS Y MODELOS CON REGRESORES ESTOCÁSTICOS. VARIABLES INSTRUMENTALES Y M.C. EN DOS FASES

Los modelos de regresión lineal típica asumen que los errores de la variable dependiente no están correlacionados con la variable o variables independientes. Cuando esto no ocurre (por ejemplo, cuando las relaciones entre las variables son bidireccionales), la regresión lineal mediante mínimos cuadrados ordinarios (OLS) deja de proporcionar estimaciones óptimas del modelo. En estos casos debe utilizarse *regresión por mínimos cuadrados en dos fases* que utiliza variables instrumentales. La regresión por mínimos cuadrados en dos fases utiliza variables instrumentales que no estén correlacionadas con los términos de error para calcular los valores estimados de los predictores problemáticos (en la primera fase) y después utiliza dichos valores calculados para estimar un modelo de regresión lineal para la variable dependiente (la segunda fase). Dado que los valores calculados se basan en variables que no están correlacionadas con los errores, los resultados del modelo en dos fases son óptimos. Suelen utilizarse como instrumentos retardos de las variables.

Como ejemplo podríamos investigar si está relacionada la demanda de un artículo (medida por el consumo) con su precio y con los ingresos del consumidor (datos en el fichero *Regresion1.sav*) . La dificultad de este modelo radica en que el precio y la demanda tienen efectos recíprocos el uno sobre el otro. Es decir, el precio puede influir en la demanda y la demanda también puede influir en el precio. Un modelo de regresión por mínimos cuadrados en dos fases permite utilizar los ingresos de los consumidores y el precio retardado (en logaritmos) para calcular un predictor sustituto del precio, el cual no esté correlacionado con los errores de medida de la demanda. Se reemplaza el precio en el modelo especificado originariamente por este sustituto y después se estima el nuevo modelo. La función de consumo que vamos a utilizar es $C_t = \beta_1 + \beta_2 Y_t + \beta_3 \log P_{t-1} + u_t$. Los datos son los mismos del apartado anterior. El modelo en términos de variables reales per cápita tiene la ecuación $V1 = \beta_1 + \beta_2 V4 + \beta_3 V6 + u$.

Para obtener un análisis de regresión por mínimos cuadrados en dos fases, elija en los menús *Analizar* → *Regresión* → *Mínimos cuadrados en dos fases* (Figura 4-87). En la Figura 4-88 seleccione la variable dependiente $V1$, seleccione una o más variables explicativas predictoras ($V4$ y $V6$) y seleccione una o más variables instrumentales ($V4$ y $V6$). Las variables explicativas no especificadas como instrumentales se consideran endógenas. Normalmente, todas las variables exógenas de la lista *Explicativas* se especifican también como variables instrumentales. El botón *Opciones* permite guardar variables y mostrar covarianzas de parámetros (Figura 4-89).

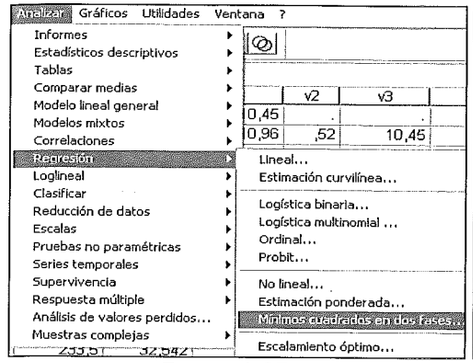


Figura 4-87

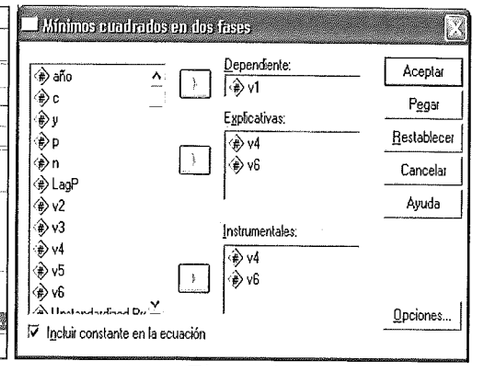


Figura 4-88

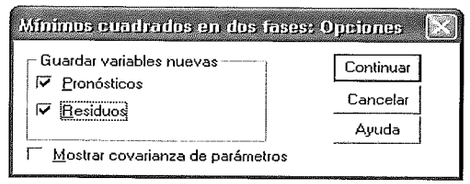


Figura 4-89

Al pulsar en *Aceptar* se obtiene la salida del procedimiento. En la Figura 4-90 se observa el principio de la salida.

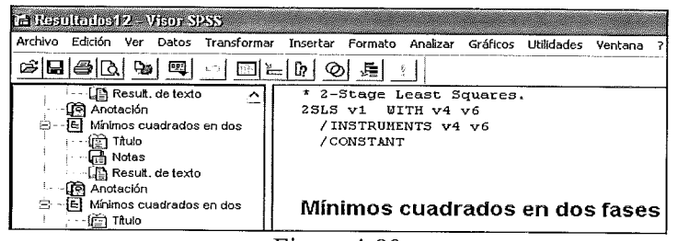


Figura 4-90

A continuación se presenta la salida completa.

Descripción del modelo

		Tipo de variable
Ecuación 1	v1	dependiente
	v4	predictor e instrumental
	v6	predictor e instrumental

MOD_1

Resumen del modelo

Ecuación 1	R múltiple	,997
	R cuadrado	,995
	R cuadrado corregida	,995
	Error típico de la estimación	,392

Figura 4-91

ANOVA

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
Ecuación 1	Regresión	845,658	2	422,829	2755,490	,000
	Residuo	4,297	28	,153		
	Total	849,954	30			

Coeficientes

		Coeficientes no estandarizados		Beta	t	Sig.
		B	Error típico			
Ecuación 1	(Constante)	-4,407	,527		-8,358	,000
	v4	,894	,029	,829	30,751	,000
	v6	1,095	,156	,189	7,022	,000

Figura 4-92

Se observa que los resultados del ajuste son muy buenos. La significatividad individual y conjunta es muy alta y el coeficiente de determinación ajustado también es muy alto. La ecuación del modelo estimado es:

$$C_t = -4{,}406776 + 0{,}894344\, Y_t + 1{,}09525\, \log P_{t-1} + u_t.$$

Si adicionalmente introducimos como instrumento un retardo de la función de consumo nos quedaría el modelo $C_t = \beta_1 + \beta_2\, C_{t-1} + \beta_3 Y_t + \beta_4 \log P_{t-1} + u_t$, que en términos reales per cápita tiene la ecuación $V1 = \beta_1 + \beta_2\, V3 + \beta_3 V4 + \beta_4\, V6 + u$.

Para ajustar este nuevo modelo, en la pantalla de entrada del procedimiento regresión en dos fases seleccione la variable dependiente V1, seleccione una o más variables explicativas predictoras (*V*3, *V*4 y *V*6) y seleccione una o más variables instrumentales (*V*3, *V*4 y *V*6), rellenando la pantalla de entrada del procedimiento de mínimos cuadrados en dos fases como se indica en la Figura 4-93.

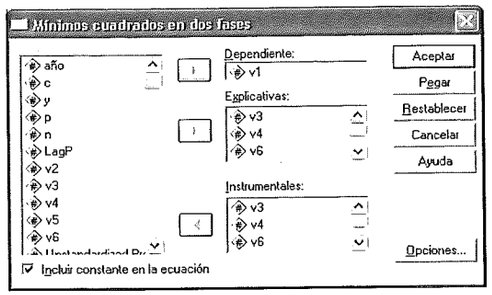


Figura 4-93

Al hacer clic en *Aceptar* se obtienen los resultados del nuevo ajuste (Figuras 4-94 y 4-45).

Resumen del modelo

Ecuación 1	R múltiple	,999
	R cuadrado	,997
	R cuadrado corregida	,997
	Error típico de la estimación	,288

Figura 4-94

ANOVA

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
Ecuación 1	Regresión	847,713	3	282,571	3404,802	,000
	Residuo	2,241	27	,083		
	Total	849,954	30			

Coeficientes

		Coeficientes no estandarizados		Beta	t	Sig.
		B	Error típico			
Ecuación 1	(Constante)	-1,849	,644		-2,871	,008
	v4	,606	,062	,561	9,797	,000
	v6	,425	,177	,073	2,400	,024
	v3	,371	,075	,376	4,977	,000

Figura 4-95

Se observa que ahora los resultados del ajuste también son muy buenos. La significatividad individual y conjunta es muy alta (superior al 97% para el peor de los parámetros) y el coeficiente de determinación ajustado también es muy alto. La ecuación del modelo estimado es:

$$V1 = -1{,}848614 + 0{,}370928\, V3 + 0{,}605666\, V4 + 0{,}424614\, V6 + u_t$$

o lo que es lo mismo:

$$C_t = -1{,}848614 + 0{,}370928\, C_{t-1} + 0{,}605666\, Y_t + 0{,}424614\, \log P_{t-1} + u_t$$

Si elegimos *Gráficos → Cuadros de diálogo antiguos→ Dispersión/Puntos → Dispersión simple* y rellenamos la pantalla de entrada con los años en el eje X y los residuos (previamente guardados en una variable en el *Editor de datos* según lo indicado en la pantalla *Opciones* de la Figura 4-89) en el eje Y (Figura 4-96) obtenemos el gráfico de la Figura 4-97, que muestra la aleatoriedad de los residuos. Este hecho indica la ausencia de autocorrelación.

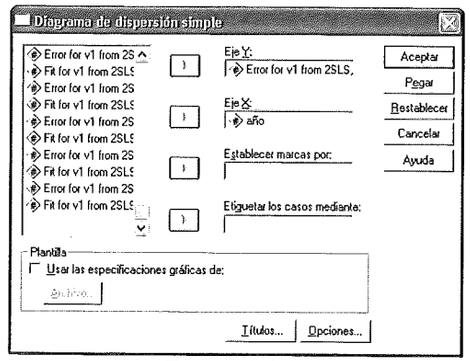


Figura 4-96

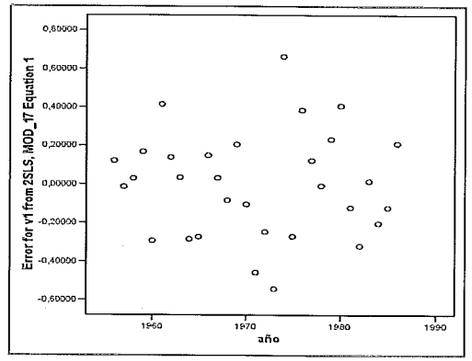


Figura 4-97

Si elegimos *Gráficos* → *Cuadros de diálogo antiguos*→ *Dispersión/Puntos* → *Dispersión simple* y rellenamos la pantalla de entrada con los pronósticos (previamente guardados en una variable en el *Editor de datos* según lo indicado en la pantalla *Opciones* de la Figura 4-89) en el eje X y los residuos en el eje Y (Figura 4-98) obtenemos el gráfico de la Figura 4-99, que muestra una distribución aleatoria de los puntos, lo que indica que no habrá problemas de linealidad ni heteroscedasticidad en el modelo.

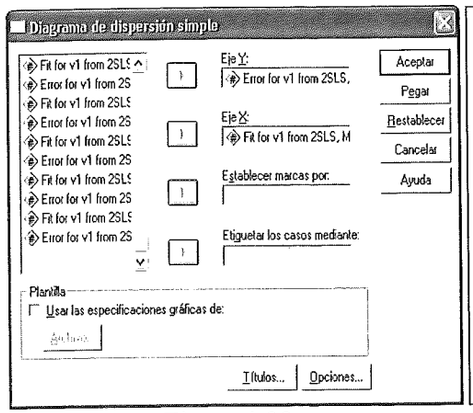


Figura 4-98

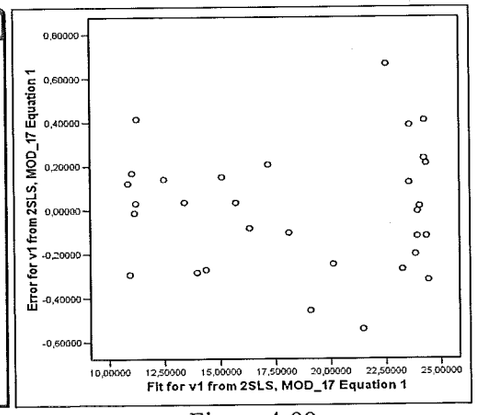


Figura 4-99

Si elegimos *Gráficos* → *Cuadros de diálogo antiguos*→ *Histograma* y tomamos como variable los residuos (Figura 4-100) obtenemos el gráfico de la Figura 4-101, que muestra que los residuos se ajustan a una distribución normal, lo que prueba que el modelo cumple la hipótesis de normalidad.

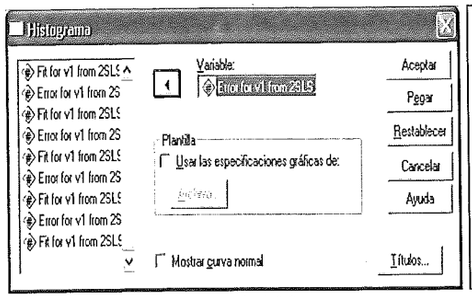


Figura 4-100

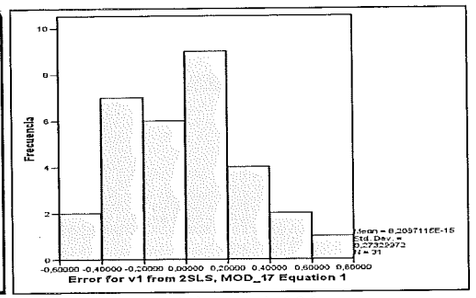


Figura 4-101

Se observa que hemos obtenido un ajuste muy aceptable del modelo completo de la función de consumo utilizando variables instrumentales y el método de estimación de los mínimos cuadrados en dos fases.

SPSS Y MODELOS CON HETEROSCEDASTICIDAD Y MULTICOLINEALIDAD. MÍNIMOS CUADRADOS PONDERADOS

La hipótesis de homoscedasticidad en un modelo exige que el término de error sea una variable aleatoria con esperanza nula y matriz de varianzas covarianzas escalar y diagonal. Es decir, para todo t la variable u_t tiene media cero y varianza σ^2 no dependiente de t, y además $Cov(u_i, u_j) = 0$ para todo i y para todo j distintos entre sí. Pero este supuesto de varianza constante no es siempre fácil de mantener.

Cuando no se da la homoscedasticidad se dice que el modelo es heteroscedástico, en cuyo caso, la regresión lineal mediante mínimos cuadrados ordinarios (MCO, OLS) deja de proporcionar estimaciones óptimas para el modelo (existe heteroscedasticidad). Si las diferencias de variabilidad se pueden pronosticar a partir de otra variable, el procedimiento *Estimación ponderada* de SPSS permite calcular los coeficientes de un modelo de regresión lineal mediante mínimos cuadrados ponderados (MCP, WLS), de forma que se les dé mayor ponderación a las observaciones más precisas (es decir, aquéllas con menos variabilidad) al determinar los coeficientes de regresión. El procedimiento *Estimación ponderada* contrasta un rango de transformaciones de ponderación e indica cuál se ajustará mejor a los datos.

Como ejemplo consideramos un modelo que explica la variable Y en función de la variable X para los datos contenidos en el fichero *ponderado.sav*. Intentaremos realizar el ajuste $Y = a + b\,X + u$.

En primer lugar intentamos ajustar nuestro modelo mediante Mínimos cuadrados Ordinarios (MCO). Para ello en SPSS se elige en los menús *Analizar* → *Regresión* → *Lineal* (Figura 4-102). En el cuadro de diálogo *Regresión lineal*, seleccione una variable numérica *dependiente* (Y) y una variable numéricas *independientes* (X) según se indica en la Figura 4-103. En el botón *Estadísticos* se hace la selección de la Figura 4-105 y en el botón Gráficos se grafica residuos contra valores predichos (Figura 4-106).

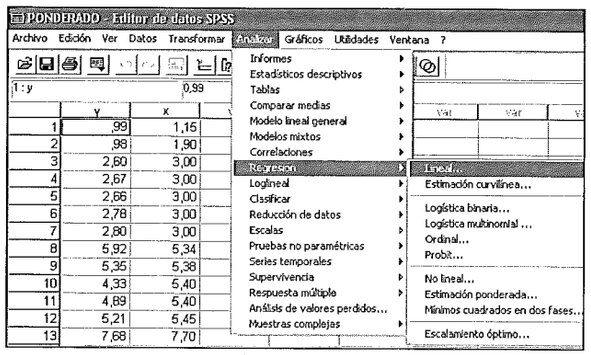


Figura 4-102

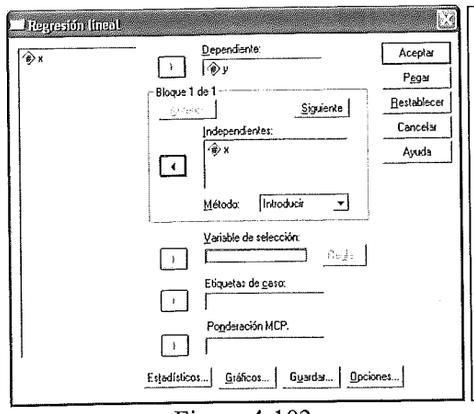


Figura 4-103

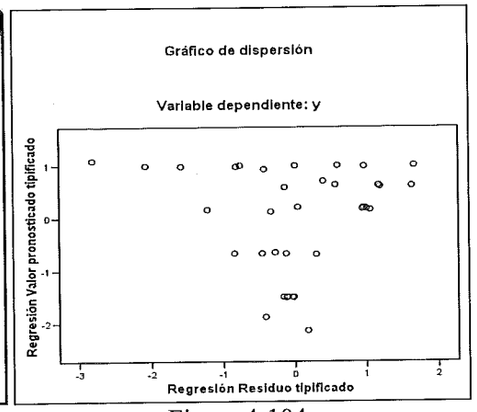


Figura 4-104

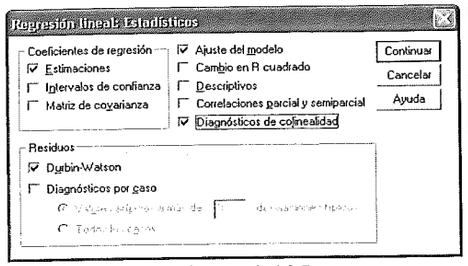


Figura 4-105

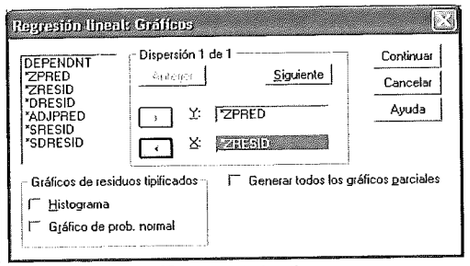


Figura 4-106

Si observamos las Figuras 4-107 y 4-108 vemos que el coeficiente de determinación ajustado (R^2) es alto, que el ajuste global es bueno (p-valor del contraste de la F muy pequeño) y que no hay multicolinealidad (índices de condición menores que 30 y FIV bajo). Pero la significatividad de la constante es mala y el gráfico de residuos tipificados contra valores predichos (Figura 4-104) no es aleatorio. Esto indica heteroscedasticidad.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,917[a]	,840	,835	1,45660	1,952

a. Variables predictoras: (Constante), x

b. Variable dependiente: y

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	367,948	1	367,948	173,422	,000[a]
	Residual	70,016	33	2,122		
	Total	437,964	34			

a. Variables predictoras: (Constante), x

b. Variable dependiente: y

Figura 4-107

Coeficientes[a]

Modelo		Coeficientes no estandarizados B	Coeficientes no estandarizados Error típ.	Coeficientes estandarizados Beta	t	Sig.	Estadísticos de colinealidad Tolerancia	Estadísticos de colinealidad FIV
1	(Constante)	-,579	,679		-,852	,400		
	x	1,135	,086	,917	13,169	,000	1,000	1,000

a. Variable dependiente: y

Diagnósticos de colinealidad[a]

Modelo	Dimensión	Autovalor	Indice de condición	Proporciones de la varianza (Constante)	Proporciones de la varianza x
1	1	1,932	1,000	,03	,03
	2	,068	5,329	,97	,97

a. Variable dependiente: y

Figura 4-108

Para solucionar el problema de la heteroscedasticidad en SPSS se utiliza el procedimiento *Estimación ponderada*. Para llevarlo a cabo, elija en los menús *Analizar* → *Regresión* → *Estimación ponderada* (Figura 4-109), seleccione Y como variable dependiente en la Figura 4-110, seleccione X como variable independientes y seleccione la variable fuente de heterocedasticidad como variable de ponderación (X). El botón *Opciones* (Figura 4-111) permite usar (y guardar como nueva variable) la mejor ponderación, eligiendo automáticamente la mejor potencia para la variable de ponderación permitiendo mostrar estimaciones y la tabla ANOVA. Al pulsar *Aceptar* se obtiene la salida (Figura 4-112).

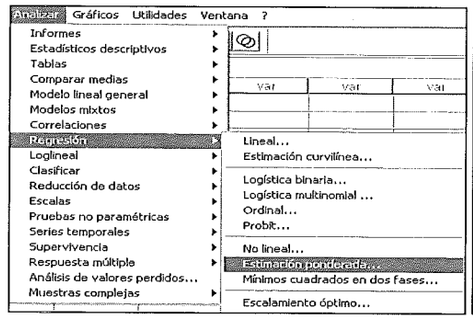


Figura 4-109

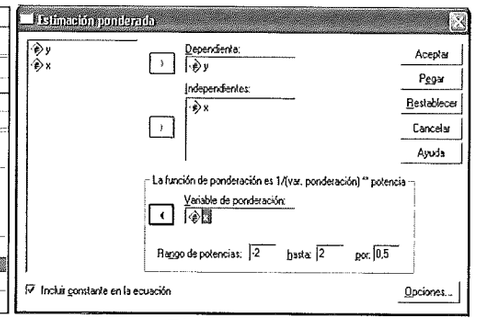


Figura 4-110

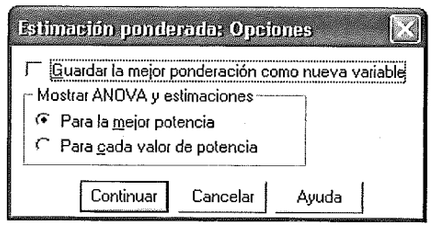


Figura 4-111

```
GET
  FILE='C:\Libros\ECONOP\ECONO2\ECONOBASICAISABEL\PONDERADO.SAV'.
DATASET NAME Conjunto_de_datos4 WINDOW=FRONT.
* Estimación ponderada.
TSET NEWVAR=NONE.
WLS y WITH x
  /SOURCE x
  /POWER -2 TO 2 BY 0.5
  /CONSTANT
  /PRINT BEST.
```

Análisis de mínimos cuadrados ponderados

[Conjunto_de_datos4] C:\Libros\ECONOP\ECONO2\ECONOBASICAISABEL\PONDERADO.SAV

Resumen de potencia

Valores de log-verosimilitud[b]

Potencia	-2,000	-74,739
	-1,500	-71,476
	-1,000	-68,224
	-,500	-64,995
	,000	-61,827
	,500	-58,797
	1,000	-56,067
	1,500	-53,932
	2,000	-52,877[a]

Figura 4-112

Los resultados del ajuste que se presentan en la Figura 4-113 indican un buen ajuste del modelo (R^2 ajustado = 0,94685) y una significatividad individual y conjunta de los parámetros estimados muy alta. La ecuación de regresión estimada será: $y = -0{,}58 + 1{,}13\,x$.

Resumen del modelo

R múltiple	,974
R cuadrado	,948
R cuadrado ajustado	,947
Error típico de la estimación	,173
Valor de la función log-verosimilitud	-52,877

ANOVA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Regresión	18,143	1	18,143	606,745	,000
Residual	,987	33	,030		
Total	19,130	34			

Coeficientes

	Coeficientes no estandarizados		Coeficientes estandarizados		t	Sig.
	B	Error típico	Beta	Error típico		
(Constante)	-,580	,190			-3,053	,004
x	1,130	,046	,974	,040	24,632	,000

Figura 4-113

SAS Y LA MULTICOLINEALIDAD, AUTOCORRELACIÓN HETEROSCEDASTICIDAD, VALORES INFLUYENTES Y ERRORES DE ESPECIFICACIÓN

SAS permite analizar los problemas de ajuste del modelo lineal a través de diferentes opciones del ***procedimiento REG***. La opción *collin* permite analizar la multicolinealidad a través de los índices de condición, la opción *DW* permite analizar la autocorrelación mediante el estadístico de Durbin Watson, las opciones *FIV* y *tol* ofrecen los factores de inflación e índices de tolerancia para el estudio de la multicolinealidad y la opción *R* presenta un análisis detallado de los residuos.

Por otra parte, el ***procedimiento AUTORREG*** estima y predice modelos de regresión con series de tiempo en presencia de autocorrelación y heteroscedasticidad y realiza contrastes de estacionariedad (raíces unitarias) y cointegración. Su sintaxis resumida es la siguiente:

PROC AUTOREG opciones;
BY variables;
MODEL dependiente = regresores / opciones;
HETERO variables / opciones;
RESTRICT ecuación , ... , ecuación;
TEST ecuación , ... , ecuación / opción;
OUTPUT OUT = conjunto de datos opciones;

Las opciones de PROC AUTORREG son DATA=conjunto de datos de entrada, OUTEST =conjunto de datos con las estimaciones y COVOUT para escribir la matriz de varianzas-covarianzas de los parámetros estimados.

La sentencia HETERO especifica las variables que están relacionadas por heteroscedasticidad y sus opciones son: LINK=EXP | SQUARE | LINEAR para especificar la forma funcional de la heteroscedasticidad, COEF=NONNEG | UNIT | ZERO | UNREST para imponer restricciones a los parámetros del modelo de heteroscedasticidad (no negativos, unitarios, nulos o sin restricción), STD=NONNEG | UNIT | UNREST para imponer restricciones a la varianza del modelo de heteroscedasticidad (no negativa, unitaria o sin restricción), TEST = LM para test de los multiplicadores de Lagrange para heteroscedasticidad, NOCONST para modelo de heteroscedasticidad sin constante.

La sentencia MODEL define el modelo a ajustar y sus opciones más importantes son: CENTER para centrar en la media la variable dependiente cuando el modelo no tiene regresores, NOINT para suprimir la constante del modelo, NLAG=orden del proceso AR del error, LDW para hallar los p-valores del estadístico de Durbin Watson por linealización aproximada, METHOD=ML | ULS | YW | ITYW define el método de estimación del modelo (máxima verosimilitud, mínimos cuadrados incondicionales, Yull Walker y Yull Walker iterado).

La sentencia RESTRICT permite introducir restricciones para el modelo definidas por las ecuaciones que se especificarán. La sentencia TEST permite realizar contrastes de funciones de los parámetros definidas por las ecuaciones que se especificarán. La sentencia OUTPUT OUT=conjunto de datos permite guardar todo tipo de información resultante en el modelo.

El ***procedimiento MODEL*** es mucho más general, pero puede utilizarse fácilmente para declarar qué variables utilizamos como instrumentos en el ajuste de un modelo dinámico con regresores estocásticos. Su sintaxis reducida es la siguiente:

PROC MODEL opciones;
PARMS parámetros;
Descripción de la ecuación;
FIT variable endógena;
INSTRUMENTS variables instrumentales;
RUN;

En la sentencia PARAMS se especifica la lista de parámetros a estimar. A continuación se especifica la ecuación a estimar. En la sentencia FIT se especifica la variable endógena del modelo. En la sentencia INSTRUMENTS se especifica el nombre de todos los instrumentos a utilizar en la estimación por variables instrumentales.

Como primer ejemplo consideramos un modelo que ajusta el consumo total en España en el periodo 1954-88 en función del PIB a precios de mercado y el consumo del año anterior. Es decir, el modelo a ajustar será:

$$CP_t = \beta_0 + \beta_1 PIBPM_t + \beta_2\, CP_{t-1} + u_t$$

Se trata de un modelo dinámico con un retardo en la variable dependiente. En primer lugar realizamos su ajuste por MCO para ver si existe autocorrelación residual. Utilizamos la siguiente sintaxis SAS:

```
data datos;
input PIBPM CP @@;
        CPret=lag1(CP);
        PIBPMret=lag1(PIBPM);
cards;
4322.2  3536.9  4546.6  3711.5  4872.5  3976.9  5080.8  4120
5309.9  4262.6  5209.2  4333.1  5331.7  4240.5  5963    4658.7
6518.2  5037.1  7089    5555.7  7527.4  5799.9  8004.1  6152.6
8568.8  6528.2  8939.1  6889.7  9544.6  7276.3  10397.9 7743.7
10822.3 8110.3  11318   8512.2  12227.1 9181    13166.9 9877.4
13866.5 10433.3 13940.9 10669.3 14397.2 11282.4 14829.2 11492
15044   11676.1 15023.1 11875   15209.1 12009.7 15171.3 11985.5
15355.9 12099.5 15633.1 12214.5 15914.4 12237.4 16282.8 12578.8
16816.3 13081.6 17748.6 13873.8 18676.5 14540.1
;
```



```
proc reg;
      model CP=PIBPM CPret/DW collin;
      output out=cons r=rest;
run;
```

La salida con el modelo ajustado por MCO con diagnosis de autocorrelación es la siguiente:

Procedimiento REG
Modelo: MODEL1
Variable dependiente: CP

Number of Observations Read	35
Number of Observations Used	34
Number of Observations with Missing Values	1

Análisis de la varianza

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Modelo	2	387397510	193698755	18823.7	<.0001
Error	31	318995	10290		
Total corregido	33	387716505			
Raíz MSE		101.44046	R-cuadrado	0.9992	
Media dependiente		8765.18824	R-Cuad Adj	0.9991	
Var Coeff		1.15731			

Estimadores del parámetro

Variable	DF	Estimador del parámetro	Error estándar	Valor t	Pr > |t|
Intercept	1	112.95624	48.64653	2.32	0.0270
PIBPM	1	0.43381	0.03990	10.87	<.0001
CPret	1	0.44404	0.05237	8.48	<.0001

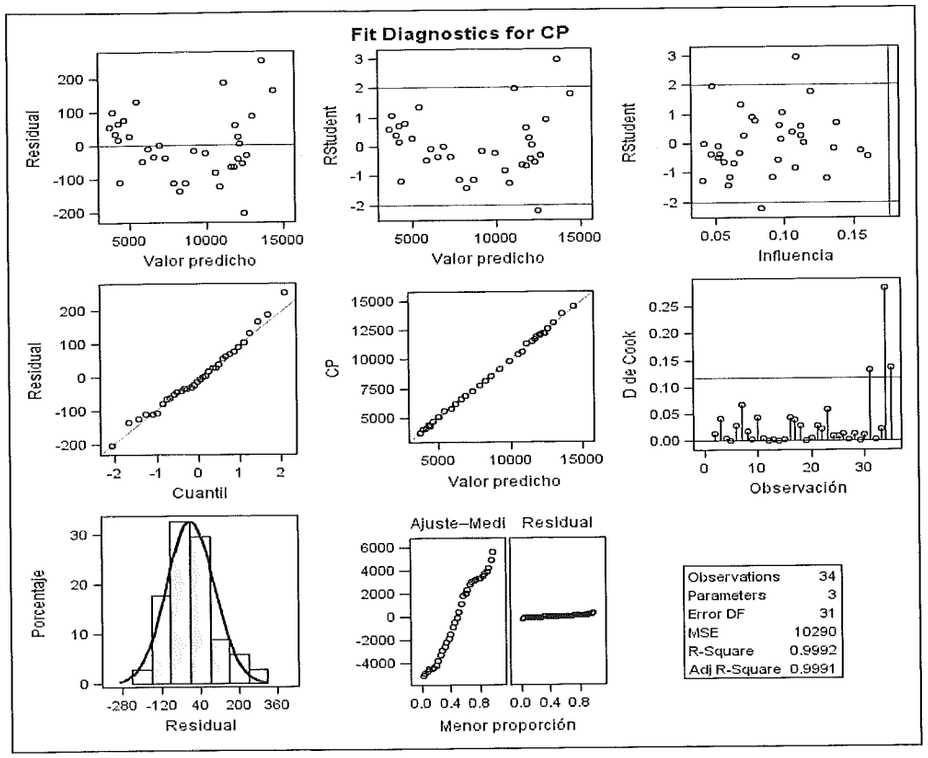
D de Durbin-Watson	1.353
Número de observaciones	34
1 Autocorrelación de orden	0.276

Se observa que el modelo ajustado resulta ser:

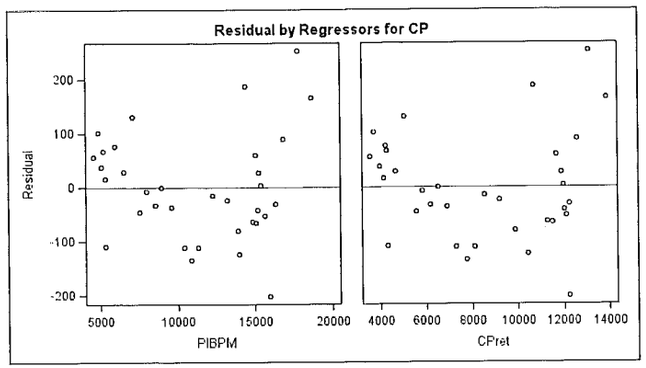
$$CP_t = 112{,}95 + 0{,}43\, PIBPM_t + 0{,}44\, CP_{t-1} + u_t$$

También se observa que la significatividad conjunta de los parámetros es muy buena (p-valor de la F de Fisher muy pequeño igual a 0,0001). La significatividad individual de cada parámetro estimado también es muy buena (p-valores de la T de Student menores que 0,05 con valores respectivos 0,027, 0,0001 y 0,0001). Además, los dos primeros índices de condición son menores que 30, lo que indica que no habrá problemas de ***multicolinealidad***.

El gráfico que contiene el panel de diagnosis residual se presenta a continuación:



Si observamos el panel de diagnosis se muestra que los gráficos de residuos contra valores predichos son más o menos aleatorios con lo que no habrá problemas de ***heteroscedasticidad*** ni de ***especificación no lineal***. También so aleatorios los gráficos de residuos contra las variables independientes del modelo que se muestran a continuación, lo que corrobora la ausencia de heteroscedasticidad y de ***errores de especificación***.



Para estudiar los ***puntos influyentes***, el panel de diagnosis muestra los gráficos de la *influencia* contra los residuos estudentizados y del número de observación contra la *Distancia de Cook*, observándose una única observación claramente influyente (se sale de las bandas de confianza en ambos gráficos). Tampoco hay problemas de ***normalidad residual***, ya que el histograma residual se ajusta bien a la campana de Gauss y el gráfico de cuantiles contra residuos se ajusta bien a la diagonal del primer cuadrante.

Pero el principal problema es el valor del estadístico DW de Durbin Watson 1,353, que se aleja demasiado de su valor ideal 2. Estamos entonces ante problemas de autocorrelación residual en el modelo. Esto hace que las estimaciones por MCO sean inconsistentes y será necesario utilizar métodos alternativos, como por ejemplo utilizar el procedimiento AUTORREG inicialmente con una estructura de un retardo para el residuo con la siguiente sintaxis:

```
proc autoreg;
      model CP=PIBPM CPret/nlag=1;
run;
```

La salida es la siguiente:

```
                    Procedimiento AUTOREG

                 Variable dependiente    CP

            Estimadores de mínimos cuadrados ordinarios

SSE              318995.162   DFE                          31
MSE                   10290   Raiz MSE               101.44046
SBC              418.050305   AIC                   413.471224
MAE              76.2035092   AICC                  414.271224
MAPE             0.95865145   HQC                   415.032822
Durbin-Watson         1.3529  R-cuadrado de regresión    0.9992
                              R-cuadrado total           0.9992

                                      Error              Aprox
Variable        DF   Estimador      estándar   Valor t   Pr > |t|

Intercept        1    112.9562       48.6465      2.32     0.0270
PIBPM            1      0.4338        0.0399     10.87     <.0001
CPret            1      0.4440        0.0524      8.48     <.0001

                 Estimadores de autocorrelaciones

Retardo    Covarianza    Correlation    -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1

   0        9382.2       1.000000      |                    |********************|
   1        2591.3       0.276193      |                    |******              |

                 MSE preliminar      8666.5
```

Estimadores de parámetros autoregresivos

Retardo	Coeficiente	Error estándar	Valor t
1	-0.276193	0.175472	-1.57

Estimadores Yule-Walker

SSE	287888.802	DFE	30
MSE	9596	Raíz MSE	97.96067
SBC	418.167559	AIC	412.062117
MAE	73.7464353	AICC	413.441428
MAPE	0.94230382	HQC	414.144248
Durbin-Watson	1.7914	R-cuadrado de regresión	0.9987
		R-cuadrado total	0.9993

Procedimiento AUTOREG

Variable	DF	Estimador	Error estándar	Valor t	Aprox Pr > |t|
Intercept	1	104.2009	62.8209	1.66	0.1076
PIBPM	1	0.4667	0.0488	9.55	<.0001
CPret	1	0.4014	0.0643	6.25	<.0001

Se observa que el modelo ajustado resulta ser ahora:

$$CP_t = 104{,}2 + 0{,}46\ PIBPM_t + 0{,}4\ CP_{t-1} + u_t - 0{,}276193\ u_{t-1}$$

También se observa que la significatividad individual de cada parámetro estimado es muy buena (p- valores de la T de Student menores que 0,05 con valores respectivos 0,1, 0,0001 y 0,0001). Además, el estadístico de Durbin-Watson vale ahora 1,8 (muy cercano al valor ideal 2) lo que indica que se ha solucionado el problema de la autocorrelación en el modelo.

Otra alternativa para solucionar el problema de autocorrelación es utilizar variables instrumentales para el ajuste. Tomamos como instrumento de la variable dependiente desfasada CP_{t-1} (Cret) el PIB desfasado una unidad temporal ($PIBPM_{t-1}$), variable que denominaremos PIBPMret. El ***ajuste con variables instrumentales*** se realiza con el proc MODEL de SAS como se indica a continuación:

```
proc model;
      parms b0 b1 b2;
      CP=b0+b1*PIBpm+b2*CPret;
      fit CP;
      instruments PIBpm PIBpmret;
run;
```

La salida es la siguiente:

The MODEL Procedure

Model Summary

Model Variables	1
Parameters	3
Equations	1
Number of Statements	1

Model Variables CP
Parameters b0 b1 b2
Equations CP

The Equation to Estimate is

CP = F(b0(1), b1(PIBpm), b2(CPret))
Instruments 1 PIBpm PIBPMret

NOTA: At 2SLS Iteration 1 convergence assumed because OBJECTIVE=1.008376E-20 is almost zero (<1E-12).

The MODEL Procedure
2SLS Estimation Summary

Data Set Options

DATA= CONS

Minimization Summary

Parameters Estimated	3
Method	Gauss
Iterations	1

Final Convergence Criteria

R	1
PPC	0
RPC(b0)	1025311
Object	1
Trace(S)	10938.96
Objective Value	1.01E-20

Observations Processed

Read	35
Solved	35
Used	34
Missing	1

The MODEL Procedure

No lineal 2SLS Resumen de errores residuales

Ecuación	Modelo DF	Error DF	SSE	MSE	Raiz MSE	R-cuadrado	R-Sq adj
CP	3	31	339108	10939.0	104.6	0.9991	0.9991

No lineal 2SLS Parameter Estimadores

Parameter	Estimador	Err std aprox	Valor t	Aprox Pr > |t|
b0	103.5565	50.2961	2.06	0.0480
b1	0.48931	0.0467	10.48	<.0001
b2	0.370823	0.0614	6.04	<.0001

Número de observaciones		Estadísticos para el sistema	
Used	34	Objective	1.008E-20
Missing	1	Objective*N	3.428E-19

Se observa que los parámetros estimados y su significatividad no difieren mucho de los obtenidos por el proc AUTORREG. El modelo ajustado con variables instrumentales resulta ser:

$$CP_t = 103{,}55 + 0{,}489\ PIBPM_t + 0{,}37\ CP_{t-1} + u_t$$

SAS Y LOS MODELOS ARCH Y GARCH

El procedimiento AUTORREG de SAS estima y predice modelos de regresión con series de tiempo en presencia de autocorrelación y heteroscedasticidad y realiza ajustes de modelos ARCH y GARCH. Su sintaxis es la siguiente:

PROC AUTOREG opciones;
  BY variables;
  MODEL dependiente = regresores / opciones;
  HETERO variables / opciones;
  RESTRICT ecuación , ... , ecuación;
  TEST ecuación , ... , ecuación / opción;
  OUTPUT OUT = conjunto de datos opciones;

Todas las sentencias se han explicado anteriormente y ahora ampliamos MODEL. La sentencia MODEL define el modelo a ajustar y sus opciones son: CENTER para centrar en la media la variable dependiente cuando el modelo no tiene regresores, NOINT para suprimir la constante del modelo, NLAG=orden del proceso AR del error, GARCH=(lista) especifica el tipo de modelo GARCH a ajustar (la lista puede ser p=a q=b para un modelo GARCH(a,b), o q=(a b) para un modelo ARCH(a,b)), BACKSTEP para eliminar del modelo los parámetros autorregresivos, SLSTAY=nivel de significación para BACKSTEP, CONVERGE=valor para la convergencia, INITIAL= valores iniciales para algunos de los parámetros estimados, LDW para hallar los p-valores del estadístico de Durbin Watson por linealización aproximada, MAXITER=número máximo de valores para la convergencia, METHOD=ML | ULS | YW | ITYW define el método de estimación del modelo (máxima verosimilitud, mínimos cuadrados incondicionales, Yull Walker y Yull Walker iterado), NOMISS para no considerar valores desaparecidos, OPTMETHOD= QN | TR para método técnica de optimización en estimaciones GARCH y un grupo de opciones de impresión que por su importancia se mencionan en la tabla siguiente:

ALL para impresión de todas las opciones	ARCHTEST imprime estadístico Q y LM para ausencia de efectos ARCH	CHOW= (obs_1 ... obs_n) para realizar el test de Chow
COEF imprime coeficientes de transformación para p observaciones	CORRB imprime correlaciones estimadas	COVB imprime covarianzas estimadas
COVEST= OP | HESSIAN | QML define tipo de matriz de covarianzas en modelos GARCH	DW= n imprime el estadístico de Durbin Watson de orden *n*	DWPROB da p-valores para el test generalizado de Durbin Watson en muestras grandes
GINV da la inversa de la matriz de Toeplitz de autocovarianzas	GODFREY da el multiplicador de Godfrey	ITPRINT para impresión en cada iteración
LAGDEP da el estadístico de Durbin Watson con variables dependientes retardadas	LAGDEP= nombre del estadístico h de Durbin con regresores retardados	LOGLIKL da el valor del log de la verosimilitud con normalidad en los errores
NOPRINT para eliminar el output	NORMAL para el test de normalidad de jarque-Bera	PARTIAL da autocorrelaciones parciales
PCHOW= (obs_1 ... obs_n) para el test de Chow	RESET da el test reset de Ramsey	STATIONARITY= (PHILLIPS) para el test de estacionariedad de Phillips Perron
URSQ de el R^2 no centrado		

Como ejemplo consideramos una serie de 120 datos de la variable *y* para ajustarla en función del tiempo *t* con residuos AR(2) y sin heteroscedasticidad.

```
data a;
input t y @@;
cards;
1 9.1309196275   2 11.018368148   3 11.831781135   4 11.56883326
5 10.349177047   6 11.055213558   7 11.716826414   8 12.155459242
9 13.24868392    10 14.095215817  11 15.42618632   12 16.892686737
13 18.545055204  14 18.772883501  15 17.035026252  16 16.302250391
17 16.213965208  18 16.986991433  19 17.623332729  20 19.1311179
21 19.785707187  22 20.50940054   23 20.756296609  24 19.792424208
25 20.428854573  26 20.211834634  27 20.721834124  28 21.911981517
29 23.273483548  30 24.839053075  31 25.378770973  32 26.809935973
33 27.631293462  34 27.242697653  35 26.942149692  36 26.520494066
37 25.419840091  38 26.870900387  39 28.73044885   40 30.047369089
41 32.158233406  42 34.921524565  43 35.931012847  44 36.094050196
45 33.991465673  46 33.054768324  47 32.750108312  48 33.952843825
49 35.931126355  50 35.863280977  51 36.434040074  52 36.334661788
53 35.04076917   54 35.33421996   55 36.040234495  56 38.376439977
57 38.555976452  58 38.970110818  59 38.359856862  60 40.311848081
61 38.48355534   62 35.108278259  63 34.763455336  64 37.458986695
65 44.857421094  66 46.083423209  67 46.883958323  68 44.41588562
69 45.127322191  70 41.369416557  71 42.803704081  72 42.517361445
73 42.201251616  74 41.701190814  75 42.214652181  76 45.058702693
77 44.628087351  78 47.313086743  79 45.778395128  80 41.599936098
81 44.793359258  82 47.930784718  83 50.151666444  84 54.166917193
85 57.89964651   86 57.051208535  87 56.762396429  88 56.190230008
89 57.705277966  90 57.264268636  91 56.99024472   92 56.80258689
93 56.872412322  94 56.927146     95 56.934381217  96 59.077142878
97 60.475360843  98 61.887049873  99 62.052867583  100 61.147321538
101 60.37454925  102 58.870750    103 59.76719902  104 59.43849833
105 60.12369038  106 59.82732430  107 60.47942289  108 61.14721144
109 60.72548868  110 62.86014535  111 64.95976274  112 66.41897850
113 67.30626017  114 66.29736159  115 67.44653333  116 69.08041789
117 69.90866871  118 70.45834393  119 69.27855258  120 68.62522215
;
```

```
proc autoreg data=a;
   model y = t / nlag=2 archtest dwprob;
   output out=r r=yresid;
run;
```

La salida es la siguiente:

```
                    The AUTOREG Procedure
                  Dependent Variable    y
             Ordinary Least Squares Estimates

SSE                 690.266011   DFE                    118
MSE                   5.84971   Root MSE           2.41862
SBC                560.070468   AIC            554.495484
Regress R-Square       0.9814   Total R-Square     0.9814
Durbin-Watson          0.4060   Pr < DW            <.0001
Pr > DW                1.0000

NOTE: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for
      testing negative autocorrelation.

                 Q and LM Tests for ARCH Disturbances

      Order          Q    Pr > Q          LM    Pr > LM

          1      37.5445   <.0001     37.0072     <.0001
          2      40.4245   <.0001     40.9189     <.0001
          3      41.0753   <.0001     42.5032     <.0001
          4      43.6893   <.0001     43.3822     <.0001
          5      55.3846   <.0001     48.2511     <.0001
          6      60.6617   <.0001     49.7799     <.0001
          7      62.9655   <.0001     52.0126     <.0001
          8      63.7202   <.0001     52.7083     <.0001
          9      64.2329   <.0001     53.2393     <.0001
         10      66.2778   <.0001     53.2407     <.0001
         11      68.1923   <.0001     53.5924     <.0001
         12      69.3725   <.0001     53.7559     <.0001

                                  Standard                 Approx
Variable        DF     Estimate      Error    t Value    Pr > |t|

Intercept        1       9.2217     0.4444      20.75      <.0001
t                1       0.5024   0.006374      78.83      <.0001

                 Estimates of Autocorrelations

Lag   Covariance    Correlation    -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1

  0       5.7522       1.000000    |                    |********************|
  1       4.5797       0.796158    |                    |****************    |
  2       2.8455       0.494680    |                    |**********          |

                  Preliminary MSE      1.8017
             Estimates of Autoregressive Parameters
                                  Standard
          Lag     Coefficient        Error    t Value

            1       -1.098824     0.085877     -12.80
            2        0.380158     0.085877       4.43

                   Yule-Walker Estimates

SSE                 213.954772   DFE                    116
MSE                   1.84444   Root MSE           1.35810
SBC                430.404909   AIC            419.254942
Regress R-Square       0.9345   Total R-Square     0.9942
Durbin-Watson          2.0884   Pr < DW            0.6525
Pr > DW                0.3475

NOTE: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for
      testing negative autocorrelation.
```



```
                                  Standard                 Approx
Variable        DF     Estimate      Error    t Value    Pr > |t|

Intercept        1       9.1800     0.8652      10.61      <.0001
t                1       0.5024     0.0123      40.69      <.0001
```

Como todos los p-valores del contraste ARCH son menores que 0,0001 existe fuerte heteroscedasticidad y sería erróneo hacer el ajuste al modelo:

$$y_t = 9{,}18+0{,}5024t+v_t \quad \text{donde} \quad v_t = e_t+0{,}98824v_{t-1}-0{,}380158v_{t-2}.$$

Como solución a este problema de heteroscedasticidad podemos probar el ajuste a un modelo GARCH(1,1) mediante la siguiente sintaxis:

```
proc autoreg data=a;
   model y = t / nlag=2 garch=(q=1,p=1) maxit=50;
run;
```

La salida es la siguiente:

```
                    The AUTOREG Procedure
                  Dependent Variable    y
             Ordinary Least Squares Estimates

SSE                 690.266011   DFE                    118
MSE                   5.84971   Root MSE           2.41862
SBC                560.070468   AIC            554.495484
Regress R-Square       0.9814   Total R-Square     0.9814
Durbin-Watson          0.4060

                                  Standard                 Approx
Variable        DF     Estimate      Error    t Value    Pr > |t|

Intercept        1       9.2217     0.4444      20.75      <.0001
t                1       0.5024   0.006374      78.83      <.0001

                 Estimates of Autocorrelations

Lag   Covariance    Correlation    -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1

  0       5.7522       1.000000    |                    |********************|
  1       4.5797       0.796158    |                    |****************    |
  2       2.8455       0.494680    |                    |**********          |

                  Preliminary MSE      1.8017
             Estimates of Autoregressive Parameters
                                  Standard
          Lag     Coefficient        Error    t Value

            1       -1.098824     0.085877     -12.80
            2        0.380158     0.085877       4.43

                     GARCH Estimates

SSE                 218.860967   Observations       120
MSE                   1.82384   Uncond Var   1.6299654
Log Likelihood     -187.44013   Total R-Square    0.9941
SBC                408.392696   AIC        388.880254
Normality Test         0.0839   Pr > ChiSq        0.9589

                                  Standard                 Approx
Variable        DF     Estimate      Error    t Value    Pr > |t|

Intercept        1       8.9301     0.7235      12.34      <.0001
t                1       0.5075     0.0107      47.30      <.0001
AR1              1      -1.2301     0.1078     -11.41      <.0001
AR2              1       0.5023     0.1057       4.75      <.0001
ARCH0            1       0.0850     0.0757       1.12      0.2614
ARCH1            1       0.2103     0.0847       2.48      0.0130
GARCH1           1       0.7376     0.0960       7.68      <.0001
```

Salvo el término ARCH0 todos los parámetros resultan significativos al 95%. Podemos realizar el ajuste al siguiente modelo:

$$y_t = 8{,}9301+0{,}5075t+v_t \quad \text{donde} \quad v_t = \varepsilon_t+1{,}2301v_{t-1}-0{,}5023v_{t-2}. \text{ siendo:}$$

$$\varepsilon_t = e_t\sqrt{h_t} \quad \text{con } h_t = 0{,}0850+0{,}2103\varepsilon^2_{t-1}+0{,}7376h_{t-1}$$

STATA Y LA MULTICOLINEALIDAD, AUTOCORRELACIÓN HETEROSCEDASTICIDAD, ERRORES DE ESPECIFICACIÓN Y OBSERVACIONES INFLUYENTES

El software STATA permite realizar gráficos de residuos contra valores ajustados de la variable dependiente (*rvfplot*) o contra los valores de la propia variable predictora (*rvpplot*) con vistas a la detección de problemas de autocorrelación, heteroscedasticidad o no linealidad en los modelos econométricos. También permite realizar gráficos de dispersión entre las propias variables del modelo (*avplot* y *avplots*) con la finalidad de observar posibles problemas de multicolinealidad.

Como ejemplo consideramos las variables del fichero *auto.dta* y ajustamos una regresión que explique el precio de los automóviles en función de su peso, consumo, origen y la intersección del origen y el peso.

```
. use auto
(1978 Automobile Data)

. generate forXmpg=foreign*mpg

. regress price weight mpg forXmpg foreign
      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  4,    69) =   21.22
       Model |  350319665     4  87579916.3           Prob > F      =  0.0000
    Residual |  284745731    69  4126749.72           R-squared     =  0.5516
-------------+------------------------------           Adj R-squared =  0.5256
       Total |  635065396    73  8699525.97           Root MSE      =  2031.4

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   4.613589   .7254961     6.36   0.000     3.166263    6.060914
         mpg |   263.1875   110.7961     2.38   0.020     42.15527    484.2197
     forXmpg |  -307.2166   108.5307    -2.83   0.006    -523.7294   -90.70368
     foreign |   11240.33   2751.681     4.08   0.000     5750.878    16729.78
       _cons |  -14449.58    4425.72    -3.26   0.002    -23278.65    -5620.51
------------------------------------------------------------------------------
```

Se observa un buen ajuste del modelo con significatividad individual y conjunta del modelo muy buena. Quizá R^2 sea un poco bajo.

A continuación realizamos el gráfico de residuos contra valores predichos que incluya la línea horizontal en el origen (Figura 4-114).


```
. rvfplot, yline(0)
```

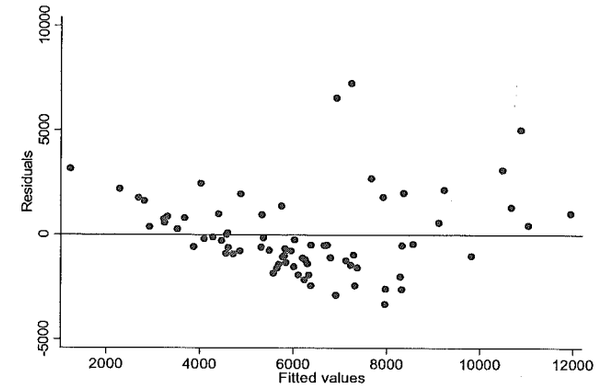


Figura 4-114

Se observa que la estructura de los puntos del gráfico no es muy aleatoria presentando un patrón de curvatura que insinúa que la relación no es lineal. Asimismo, se observa un patrón creciente de los datos hacia la derecha de la gráfica, lo que podría llevarnos a la sospecha de existencia de heteroscedasticidad.

También podemos ver el gráfico de dispersión de cada regresor contra la variable dependiente (Figura 4-115). Esta gráfica puede utilizarse para analizar las relaciones entre las variables del modelo con la variable dependiente.

```
. avplots
```

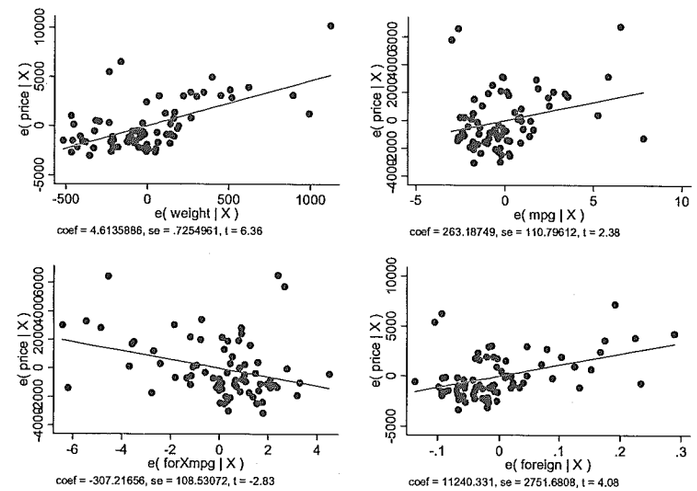


Figura 4-115

A continuación realizamos la gráfica de los residuos contra los valores de la variable independiente *mpg* (Figura 4-116).

Esta gráfica debe presentar una estructura aleatoria de sus puntos.Si para alguna variable independiente, esta gráfica no es aleatoria, dicha variable es un candidato a ser culpable de introducir heteroscedasticidad en el modelo.

```
. rvpplot mpg
```

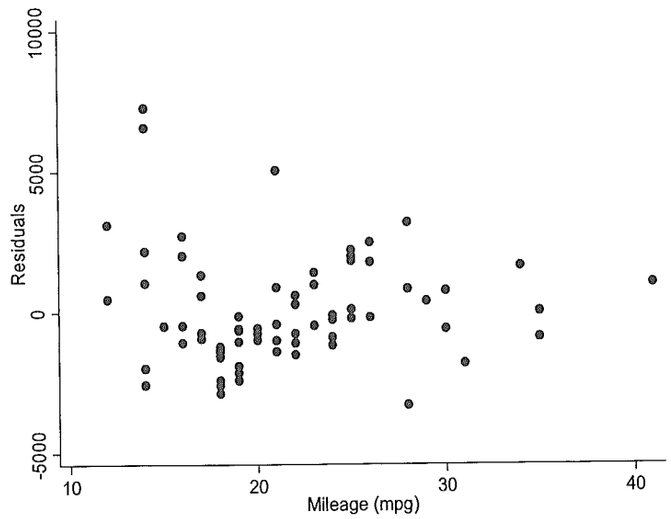


Figura 4-116

El comando *predict* se utiliza para calcular valores predichos de la variable dependiente (opción *xb*), residuos (opción *residuals*), residuos estandarizados (*rstandard*), residuos estudentizados (*rstudent*), estadísticos de influencia (*leverage, cooksd, dfbeta, dfits, welsch* y *covratio*), errores estándar de la predicción (*stdp y stdf*) y errores estándar de los residuos (*stdr*).

A continuación se analizan las observaciones influyentes mediante los estadísticos *leverage* y *dfits* para el modelo ajustado anteriormente y se guardan en las variables *xdist* y *zdist* respectivamente.

```
. predict xdist, leverage

. summarize xdist, detail
```



```
                              Leverage
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0192852       .0192852
 5%     .0198324       .0195139
10%     .0217658       .0198281       Obs                  74
25%     .0266526       .0198324       Sum of Wgt.          74

50%     .0491565                      Mean           .0675676
                        Largest       Std. Dev.      .0587061
75%     .0784069       .2030623
90%     .1439921       .2059959       Variance       .0034464
95%     .2030623       .2212826       Skewness       2.411152
99%     .3581521       .3581521       Kurtosis       10.38242

. predict zdist, dfits

. summarize zdist, detail

                                Dfits
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -.9384231      -.9384231
 5%    -.2974069      -.4402165
10%    -.2110095      -.2984942       Obs                  74
25%    -.1222725      -.2974069       Sum of Wgt.          74

50%    -.0499477                      Mean           .0364329
                        Largest       Std. Dev.      .3119636
75%     .1387799       .5724172
90%     .3069244       .8760136       Variance       .0973213
95%     .5724172       .9564455       Skewness       1.330149
99%     1.356619       1.356619       Kurtosis       8.061579
```

Suele ser conveniente disponer de los residuos, los residuos estandarizados y los residuos estudentizados almacenados en variables (*residuos, esta* y *estu*).

```
. predict residuos, residuals
. predict esta, rstandar
. predict estu, rstudent
. list residuos esta estu
```

	residuos	esta	estu
1.	-759.3586	-.3785332	-.376171
2.	-731.1284	-.3673463	-.3650317
3.	278.5821	.1402231	.1392231
4.	-992.332	-.493324	-.490602
5.	-494.6731	-.2488081	-.2471095
6.	-1431.664	-.7133379	-.7107756
7.	1771.403	.9010705	.8998269
8.	-757.7396	-.3766557	-.3743014
9.	2709.857	1.35628	1.364731
10.	-2155.183	-1.071628	-1.072799
11.	2173.117	1.102038	1.103781
12.	7271.96	3.677172	4.071051
13.	5036.348	2.74021	2.881586
14.	381.4716	.1971406	.195762
15.	-1080.561	-.5411388	-.5383468

STATA también dispone de comandos directos para detectar los problemas de variables omitidas en un modelo mediante el test de Ramsey (*estat ovtest*) y heteroscedasticidad mediante el test de Breush Pagan (*estat hettest*), el test del rango de Szroeter (*estat szroeter*), el test de White (*estat imtest*) y los factores de inflación (*estat vif*). Los problemas de autocorrelación pueden detectarse mediante el test ARCH (*estat archlm*), el test de Breusck Godfrey (*estat bgodfrey*), el test de Durbin Watson (*estat dwatson*) y el test alternativo para correlación serial de Durbin Watson (*estat dwatson*).

Más adelante veremos cómo todos estos contrastes están disponibles en Stata también a través de menús.

A continuación analizamos los problemas de variables omitidas y heterocedasticidad en nuestro modelo.

```
. estat ovtest

Ramsey RESET test using powers of the fitted values of price
       Ho:  model has no omitted variables

                 F(3, 66) =      7.77
                 Prob > F =      0.0002

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of price

         chi2(1)      =     6.50
         Prob > chi2  =   0.0108

. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      18.86     10    0.0420
            Skewness |      11.69      4    0.0198
            Kurtosis |       2.33      1    0.1273
---------------------+-----------------------------
               Total |      32.87     15    0.0049
---------------------------------------------------
```

A continuación listamos los factores de inflación (resultan demasiado altos).

```
. estat vif
```



```
    Variable |       VIF       1/VIF
-------------+----------------------
     forXmpg |     29.70    0.033671
     foreign |     28.36    0.035255
         mpg |      7.27    0.137580
      weight |      5.62    0.177810
-------------+----------------------
    Mean VIF |     17.74
```

Según los p-valores de los contrastes, pueden existir problemas de variables omitidas, heteroscedasticidaad y asimetría no normal en el modelo.

Para analizar la autocorrelación consideramos el modelo de Klein que explica el consumo en función del salario y utilizamos los contrastes habituales.

```
. use klein

. tsset yr
        time variable:  yr, 1920 to 1941

. regress consump wagegovt

      Source |       SS       df       MS              Number of obs =      22
-------------+------------------------------           F(  1,    20) =   17.72
       Model |  532.567711     1  532.567711           Prob > F      =  0.0004
    Residual |  601.207167    20  30.0603584           R-squared     =  0.4697
-------------+------------------------------           Adj R-squared =  0.4432
       Total |  1133.77488    21  53.9892799           Root MSE      =  5.4827

------------------------------------------------------------------------------
     consump |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    wagegovt |    2.50744   .5957173     4.21   0.000     1.264796    3.750085
       _cons |   40.84699   3.192183    12.80   0.000     34.18821    47.50577
------------------------------------------------------------------------------

. estat dwatson

Durbin-Watson d-statistic(  2,    22) =  .3217998

. estat durbinalt

Durbin's alternative test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |         35.035               1                   0.0000
---------------------------------------------------------------------------
                        H0: no serial correlation

. estat bgodfrey
```

```
Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |        14.264               1                   0.0002
---------------------------------------------------------------------------
                        H0: no serial correlation

. estat archlm, lags(1 2 3)
LM test for autoregressive conditional heteroskedasticity (ARCH)

---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |         5.543               1                   0.0186
       2     |         9.431               2                   0.0090
       3     |         9.039               3                   0.0288
---------------------------------------------------------------------------
          H0: no ARCH effects        vs.   H1: ARCH(p) disturbance
```

Todos los contrastes rechazan la hipótesis nula de no autocorrelación. Por tanto existen problemas de autocorrelación en este modelo. El test ARCH rechaza la hipótesis nula de no existencia de heteroscedasticidad condicional para los mordenes 1, 2 y 3.

STATA Y LA MULTICOLINEALIDAD, AUTOCORRELACIÓN, HETEROSCEDASTICIDAD, ERRORES DE ESPECIFICACIÓN Y OBSERVACIONES INFLUYENTES A TRAVÉS DE MENÚS

Después de realizar la regresión siguiente:

```
. use auto
(1978 Automobile Data)

. generate forXmpg=foreign*mpg

. regress price weight mpg forXmpg foreign
```

que también podría haberse hecho a través de menú tal y como hemos visto en un capítulo anterior, vamos a ejecutar varios contrastes de diagnosis de la regresión utilizando la ruta *Statistics → Linear model and related → Regression diagnostics → Specification test, etc*, (Figura 4-117) y eligiendo el test adecuado en la Figura 4-118. El tratamiento de las observaciones influyentes se realiza a través de las opciones *DFBETAs* y *Leverage-versus Squared-residuals plots* de la Figura 4-117.

En la Figura 4-118 hemos elegido el test de Breus Pagan de heteroscedasticidad, cuyo p-valor pequeño permite rechazar la hipótesis nula de varianza constante, existiendo por tanto heteroscedasticidad.

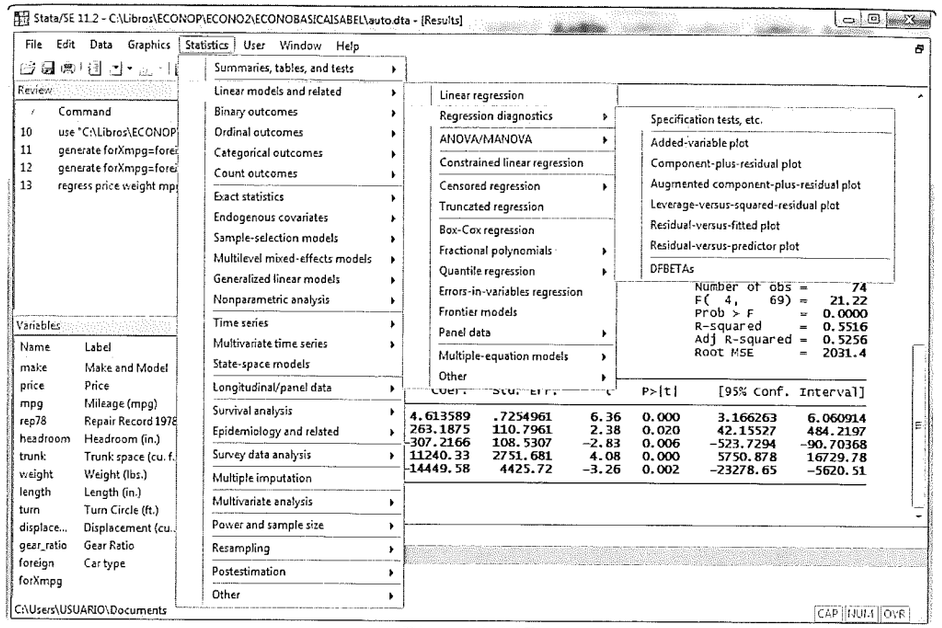


Figura 4-117

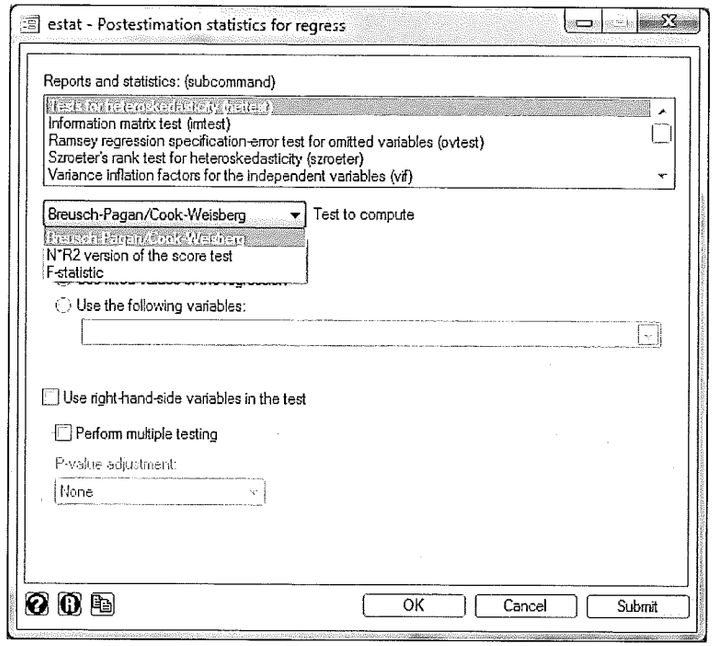


Figura 4-118

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of price

         chi2(1)      =     6.50
         Prob > chi2  =   0.0108
```

En la Figura 4-119 hemos elegido el test de Ramsey de especificación del modelo, cuyo p-valor pequeño permite rechazar la hipótesis nula de que el modelo no presenta variables omitidas. Por lo tanto existen problemas de especificación del modelo.

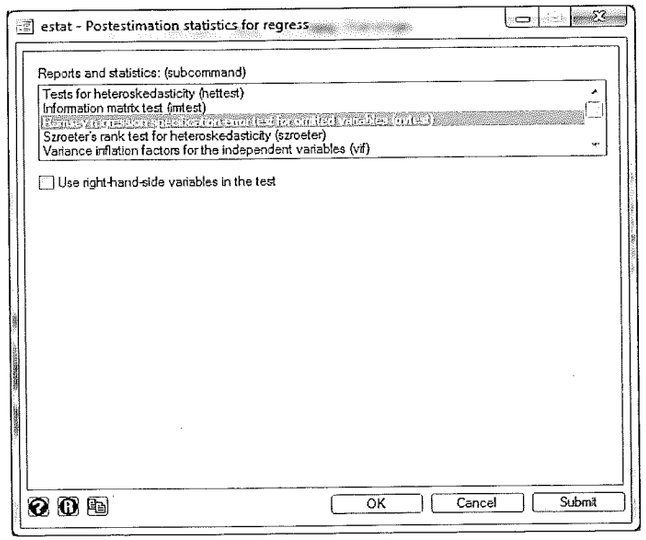


Figura 4-119

```
. estat ovtest

Ramsey RESET test using powers of the fitted values of price
       Ho:  model has no omitted variables
                 F(3, 66)  =      7.77
                 Prob > F  =    0.0002
```

En la Figura 4-120 hemos elegido la obtención de los factores de inflación de la varianza VIF para detectar la multicolinealidad, cuyos valores elevados (mayores que 10) para las variables *foreing* y *forXmpg* nos llevan a la sospecha de existencia de multicolinealidad provocada por estas dos variables. En la Figuras 4-121 y 4-122 se realizan los contrastes de Durbin-Watson y de Breus Godfrey para la autocorrelación, cuyos p-valores altos (0,6971 y 0,6850) permite aceptar con garantía la hipótesis nula de no autocorrelación. Previamente a estos contrastes es necesario dotar a la variable dependiente del modelo de estructura de serie temporal mediante la sintaxis `tsset price`.

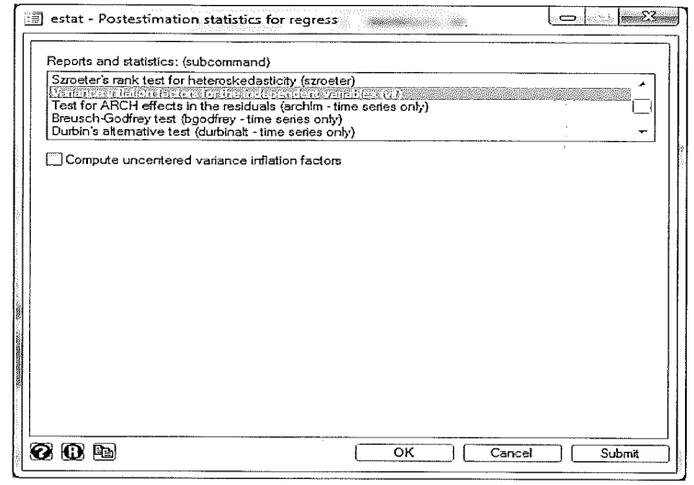


Figura 4-120

```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
     forXmpg |     29.70    0.033671
     foreign |     28.36    0.035255
         mpg |      7.27    0.137580
      weight |      5.62    0.177810
-------------+----------------------
    Mean VIF |     17.74
```

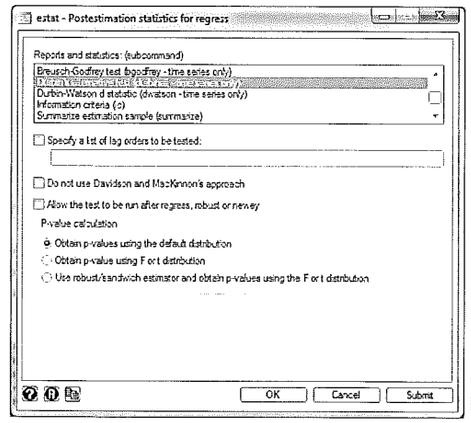


Figura 4-121

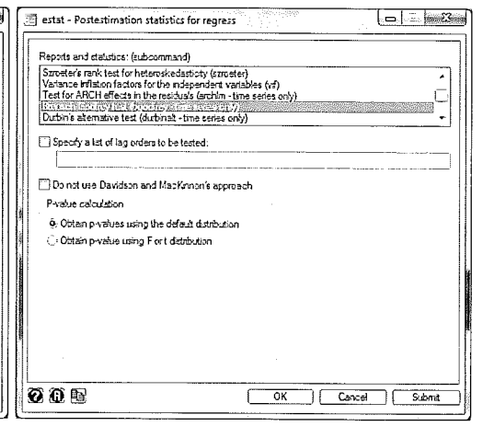


Figura 4-122

```
. tsset price
        time variable:  price, 3,291 to 15,906, but with gaps
                delta:  1 unit
```

```
. estat durbinalt

Number of gaps in sample:  69

Durbin's alternative test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |         0.152                1                   0.6971
---------------------------------------------------------------------------
                        H0: no serial correlation

. estat bgodfrey

Number of gaps in sample:  69

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |         0.165                1                   0.6850
---------------------------------------------------------------------------
                        H0: no serial correlation
```

Ejercicio 4-1. El archivo 4-1.sav contiene información correspondiente al periodo 1964–1980 de la economía española, sobre tres variables macroeconómicas: importaciones energéticas reales (IMPEN), producto interior bruto a precios de mercado (PIB), ambas en miles de millones de pesetas constantes de 1970, y el precio relativo de las importaciones energéticas (PREN).

a) Estimar el modelo de regresión que pretende explicar el volumen de importaciones energéticas en función de las variaciones del PIB y el precio relativo de las importaciones.
b) Criticar el modelo y analizar los residuos atípicos y las observaciones influyentes. Hallar intervalos de confianza para los coeficientes estimados al 95%.
c) Reestimar y criticar el modelo incluyendo una variable ficticia, con valor 0 antes de 1977 y 1 a partir de ese año.
d) Predecir el volumen de importaciones previsto para un PIB de valor 4000 y un precio relativo de las importaciones de valor 300.

Los resultados de la estimación del modelo propuesto se obtienen en SPSS mediante *Analizar* → *Regresión* → *Lineal* rellenando el cuadro de diálogo *Regresión lineal* como se indica en la Figura 4-123. El botón *Guardar* se rellena según la Figura 4-124. El botón *Estadísticos* se rellena según la Figura 4-125 y en el botón *Gráficos* se hace la selección de la Figura 4-126. Al pulsar *Aceptar* se obtiene el ajuste del modelo.

En las Figuras 4-127 a 4-129 se observa que el modelo estimado proporciona un buen ajuste, explicando un 93% de la variación total. La hipótesis de nulidad del conjunto de parámetros se rechaza claramente al 95% en el contraste de la F (p-valor menor que 0,05). Lo mismo ocurre con la hipótesis de nulidad de cada parámetro en particular, resultando todas las variables significativas (p-valores menores que 0,05). No hay multicolinealidad (índices de condición menores que 30 y FIV bajo menor que 10).

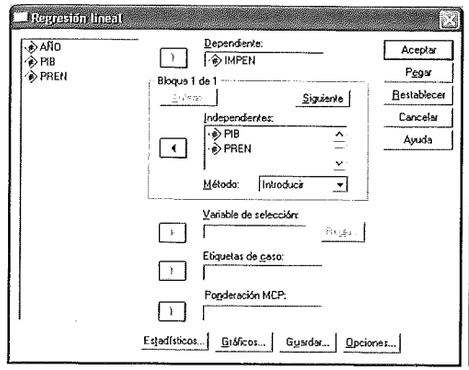


Figura 4-123

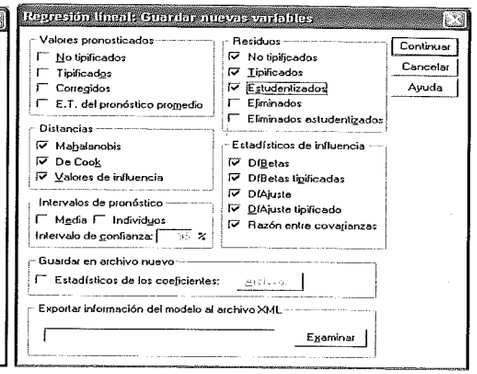


Figura 4-124

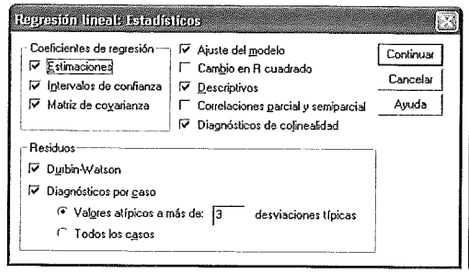


Figura 4-125

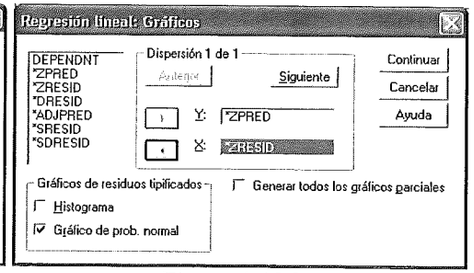


Figura 4-126

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,966[a]	,934	,927	4,00184	1,413

a. Variables predictoras: (Constante), PREN, PIB
b. Variable dependiente: IMPEN

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4281,324	2	2140,662	133,669	,000[a]
	Residual	304,279	19	16,015		
	Total	4585,604	21			

a. Variables predictoras: (Constante), PREN, PIB
b. Variable dependiente: IMPEN

Figura 4-127

Diagnósticos de colinealidad[a]

Modelo	Dimensión	Autovalor	Indice de condición	Proporciones de la varianza		
				(Constante)	PIB	PREN
1	1	2,846	1,000	,00	,00	,01
	2	,144	4,452	,09	,00	,37
	3	,011	16,252	,91	1,00	,62

a. Variable dependiente: IMPEN

Figura 4-128

Coeficientes(a)

	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza para B al 95%		Estadísticos de colinealidad	
	B	Error típ.	Beta			Límite inferior	Límite superior	Tole-rancia	FIV
Cons-tante	-21,028	4,669		-4,504	,000	-30,801	-11,255		
PIB	,027	,002	1,317	12,972	,000	,022	,031	,339	2,951
PREN	-,050	,011	-,483	-4,754	,000	-,073	-,028	,339	2,951

a Variable dependiente: IMPEN

Figura 4-129

Sin embargo, el valor del estadístico de *DW* es de 1,413, que está muy cerca de la zona de aceptación de la autocorrelación. Ello nos obliga a utilizar las tablas del contraste de Durban-Watson para el análisis más preciso de la autocorrelación. Los valores críticos en el contraste de la hipótesis nula de no autocorrelación de la perturbación, frente a la alternativa de que la perturbación se comporte como un modelo autorregresivo de orden 1 ($AR(1)$) son d_L = 0,914, d_U = 1,284 (al nivel de significación del 1%) y d_L = 1,147, d_U = 1,541 (al nivel de significación del 5%). No puede, por tanto, considerarse con toda seguridad que la perturbación no esté autocorrelacionada, puesto que al 5%, el estadístico queda en la zona de indeterminación del contraste. Por otro lado, la matriz de correlaciones (Figura 4-130) tiene un valor demasiado alto (PREN y PIB están muy correlacionadas).

Correlaciones

		IMPEN	PIB	PREN
Correlación de Pearson	IMPEN	1,000	,925	,588
	PIB	,925	1,000	,813
	PREN	,588	,813	1,000
Sig. (unilateral)	IMPEN	.	,000	,002
	PIB	,000	.	,000
	PREN	,002	,000	.
N	IMPEN	22	22	22
	PIB	22	22	22
	PREN	22	22	22

Figura 4-130

Estadísticos sobre los residuos[a]

	Mínimo	Máximo	Media	Desviación típ.	N
Valor pronosticado	20,0475	65,4906	49,9727	14,27840	22
Valor pronosticado tip.	-2,089	1,094	,000	1,000	22
Error típico del valor pronosticado	1,004	2,112	1,448	,301	22
Valor pronosticado corregido	20,4906	66,0762	49,9636	14,22604	22
Residuo bruto	-8,43317	7,46954	,00000	3,80651	22
Residuo tip.	-2,107	1,867	,000	,951	22
Residuo estud.	-2,257	1,928	-,011	1,012	22
Residuo eliminado	-9,67621	7,97180	-,09096	4,31709	22
Residuo eliminado estud.	-2,569	2,093	-,017	1,064	22
Dist. de Mahalanobis	,368	4,895	1,909	1,188	22
Distancia de Cook	,001	,250	,044	,056	22
Valor de influencia centrado	,018	,233	,091	,057	22

a. Variable dependiente: IMPEN

Figura 4-131

Para detectar residuos atípicos y observaciones influyentes observamos la tabla *Estadísticos sobre los residuos* (Figura 4-131), que presenta valores medios para todas las medidas muy razonables. Ello indica que por esta vía no habrá problemas en este modelo.

Para solucionar los problemas posibles de autocorrelación, puntos influyentes y de linealidad (mala especificación del modelo), introduciremos en la ecuación del modelo una variable ficticia que recoja el cambio estructural consecuencia del impacto de la crisis energética de 1977 en la economía española. Dicha variable, de nombre *ficticia*, valdrá 0 entre 1964 y 1976, y valdrá 1 entre 1977 y 1985. Al añadir esta variable (Figura 4-132) se obtiene el ajuste de las Figuras 4-133 y 4-134.

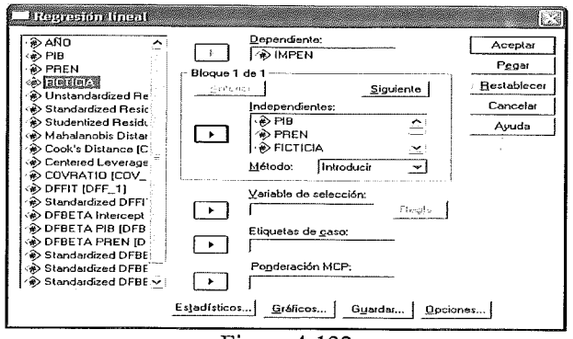


Figura 4-132

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,972[a]	,945	,939	2,95742	1,732

a. Variables predictoras: (Constante), FICTICIA, PIB, PREN

b. Variable dependiente: IMPEN

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4331,960	3	1443,987	165,096	,000[a]
	Residual	253,644	29	8,746		
	Total	4585,604	32			

a. Variables predictoras: (Constante), FICTICIA, PIB, PREN

b. Variable dependiente: IMPEN

Figura 4-133

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	-22,509	3,486		-6,457	,000		
	PIB	,027	,002	1,322	17,618	,000	,339	2,954
	PREN	-,037	,010	-,352	-3,803	,001	,222	4,495
	FICTICIA	-5,025	2,089	-,171	-2,406	,023	,377	2,652

a. Variable dependiente: IMPEN

Diagnósticos de colinealidad[a]

Modelo	Dimensión	Autovalor	Indice de condición	Proporciones de la varianza			
				(Constante)	PIB	PREN	FICTICIA
1	1	3,614	1,000	,00	,00	,00	,01
	2	,333	3,295	,02	,00	,00	,35
	3	,046	8,887	,09	,00	,60	,64
	4	,007	22,228	,89	1,00	,40	,00

a. Variable dependiente: IMPEN

Figura 4-134

El modelo ajustado será ahora:

$$IMPEN = -22{,}509 + 0{,}027 * PIB - 0{,}037 * PREN - 5{,}025 * FICTICIA$$

El nuevo modelo estimado proporciona un buen ajuste, explicando un 94,5 % de la variación total (más que el modelo inicial). La hipótesis de nulidad del conjunto de parámetros se rechaza claramente al 95% en el contraste de la F (p-valor menor que 0,05). Lo mismo ocurre con la hipótesis de nulidad de cada parámetro en particular, resultando todas las variables significativas (p-valores menores que 0,05). El valor del estadístico de DW es ahora de 1,732, que supera claramente a 1,4 y está fuera de la zona de aceptación de la autocorrelación del contraste de Durbin-Watson. Además, si representamos los residuos estandarizados contra el número de observación con *Gráficos → Cuadros de diálogo antiguos → Dispersión/Puntos → Dispersión simple* (Figura 4-135) se obtiene la Figura 4-136, que indica una distribución aleatoria de los puntos, lo que corrobora la ausencia de autocorrelación.

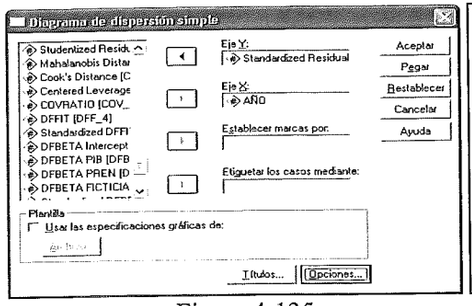


Figura 4-135

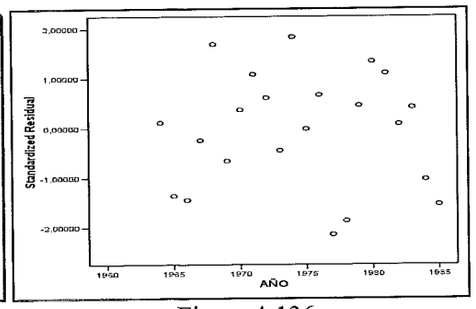


Figura 4-136

Por otro lado, el gráfico de normalidad (Figura 4-137) presenta un ajuste aceptable a la diagonal del primer cuadrante y el gráfico de residuos tipificados contra valores pronosticados tipificados (Figura 4-138) presenta una distribución aleatoria de sus puntos lo que corrobora la ausencia de heteroscedasticidad y falta de linealidad. Por otro lado, no hay multicolinealidad (índices de condición menores que 30 y FIV bajo).

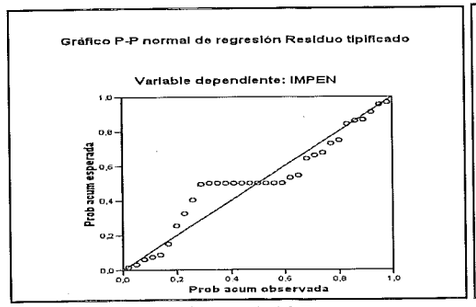


Figura 4-137

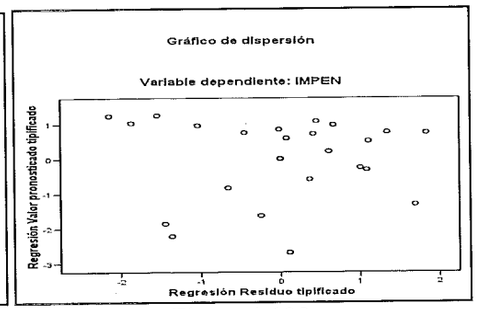


Figura 4-138

Una vez que hemos contrastado la validez del modelo ampliado con la variable ficticia, podremos utilizarlo para realizar predicciones. Por tanto el volumen de importaciones previsto para un PIB de valor 4000 y un precio relativo de las importaciones de valor 300 será -28,0805 + 0,0288368*4000 -0,0315191*300 - 9,32542*1 = 68,5.

Ejercicio 4-2. El archivo 4-2.wf1 contiene datos sobre las variables X e Y. Se quiere ajustar un modelo de regresión que tenga Y como variable dependiente y X como variable independiente, pero resulta que por errores de medición o inobservabilidad de la variable X se dispone realmente de sus datos recogidos en la variable XE. Realizar un ajuste adecuado para el modelo.

Al tratarse de un modelo con errores de medición en la variable explicativa, será necesario utilizar variables instrumentales para ajustar el modelo. Un método para obtener estimadores MCO consistentes de los parámetros, consiste en sustituir la variable explicativa sujeta a errores de medición por otra variable (denominada variable instrumental), que está muy correlacionada con la variable explicativa original y que es indepediente del término de error. La variable instrumental más sencilla que se utiliza generalmente en modelos con errores de medición en las variables suele ser la propia variable explicativa con errores retardada, en nuestro caso $XE(-1)$. Pero, para que $XE(-1)$ sea un buen instrumento de XE, es necesario probar que es independiente del término de error (endogeneidad de XE). También es necesario probar que XE y $XE(-1)$ están muy correlacionadas.

Para ***probar la endogeneidad*** de XE aplicamos el ***contraste de Hausman***. Para ello, en un primer paso planteamos la regresión de la variable XE respecto a su instrumento $XE(-1)$ con el objeto de calcular los residuos *resid01*.

Se elige *Quick → Estimate Equation*. Se rellena la pantalla *Equation Specification* de la solapa *Specification* como se indica en la Figura 4-139, se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar*. Ahora para guardar los residuos en la variable *resid01* elegimos *proc→ Make Residual Series* (Figura 4-140) y escribimos el nombre *resid01* en la figura 4-141. Al pulsar *OK*, la se obtienen los resuiltados. La fuerte significatividad individual del parámetro estimado de $XE(-1)$ en su regresión con XE indica que ambas variables están correlacionadas.

A continuación, en un segundo paso, realizamos la regresión de la variable dependiente del modelo inicial Y respecto de la variable XE y la variable *resid01*. Se elige *Quick → Estimate Equation*. Se rellena la pantalla *Equation Specification* de la solapa *Specification* como se indica en la Figura 4-142, se elige *Least Squares* en el campo *Method* (para ajustar por mínimos cuadrados) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-143 en la que se observa una significatividad cercana al 80% de la variable *resid01*. Ello nos lleva a aceptar la endogeneidad del regresor XE según el contraste de Hausman.

Una vez que sabemos que $XE(-1)$ es un buen instrumento para XE, realizamos el ajuste del modelo $Y = \beta_0 + \beta_1 XE(-1)$. Se elige *Quick → Estimate Equation*. Se rellena la pantalla *Equation Specification* de la solapa *Specification* como se indica en la Figura 4-144, se elige *Least Squares* en el campo *Method* y se hace clic en *Aceptar*. Se obtiene el ajuste de la Figura 4-145. Se observa un alto R^2 y buena significatividad individual y conjunta de los parámetros estimados.

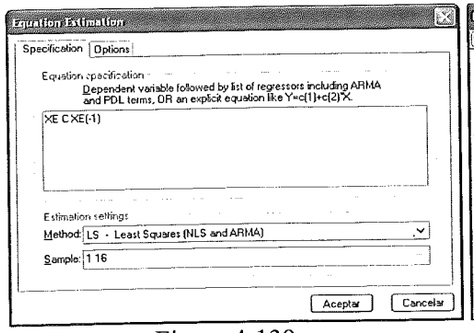


Figura 4-139

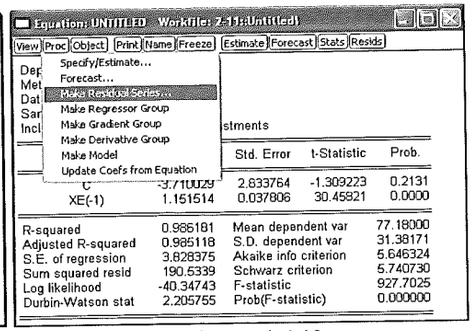


Figura 4-140

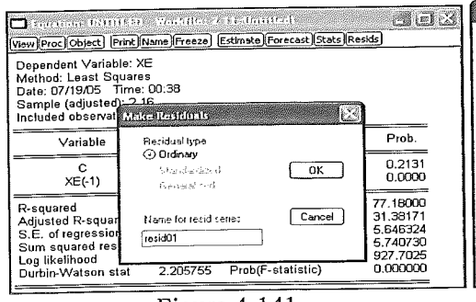


Figura 4-141

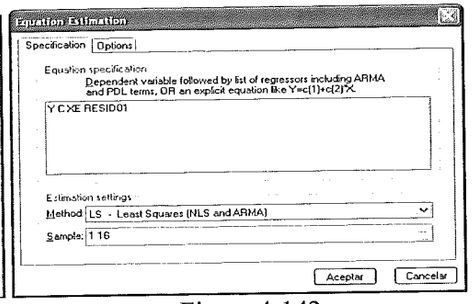


Figura 4-142

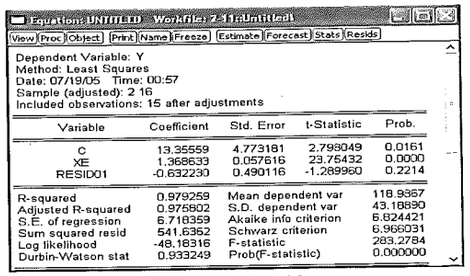


Figura 4-143

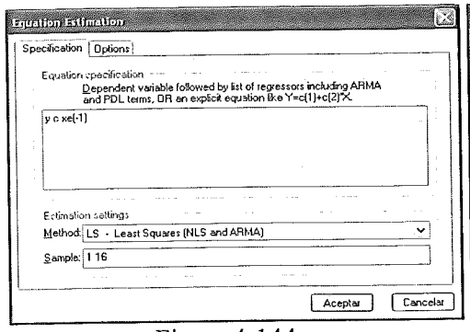


Figura 4-144

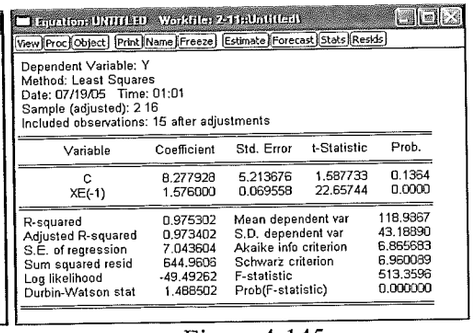


Figura 4-145

El modelo finalmente estimado es $Y = 8{,}277928 + 1{,}576XE(-1) + u$.

Pero sería más exacto realizar la estimación mediante el método de los mínimos cuadrados en dos etapas. Para ello se elige *Quick → Estimate Equation*. Se rellena la pantalla *Equation Specification* de la solapa *Specification* como se indica en la Figura 4-146 seleccionando en el campo *Method* la opción *TSLS – Two-Stage Least Squares (TSNLS and ARMA)*. En el campo *Instrument List* se introducen como instrumentos la variable exógenas y su instrumento. En el campo *Equation Specification* se introducen las variables del modelo a ajustar. Al hacer clic en *Aceptar* se obtienen los resultados de la Figura 4-147.

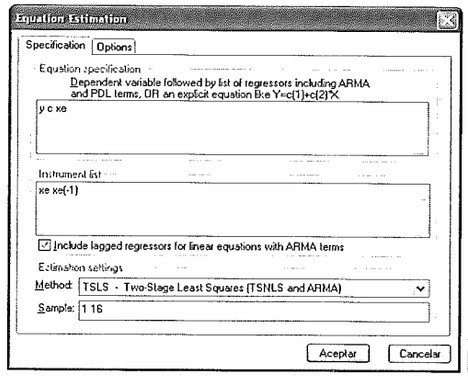


Figura 4-146

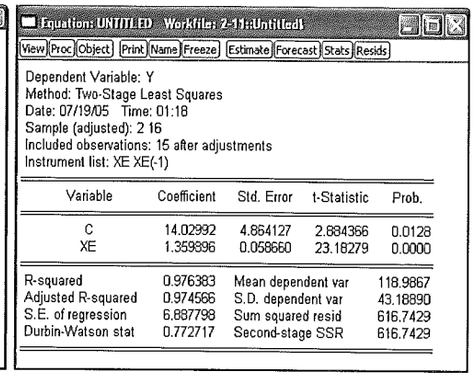


Figura 4-147

Si estimamos el modelo por el método de White para evitar la heteroscedasticidad (Figura 4-148) se obtiene el ajuste de la Figura 4-149.

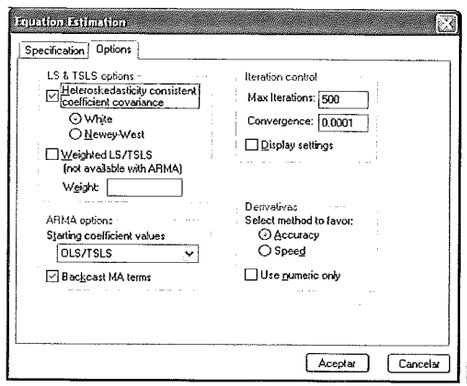


Figura 4-148

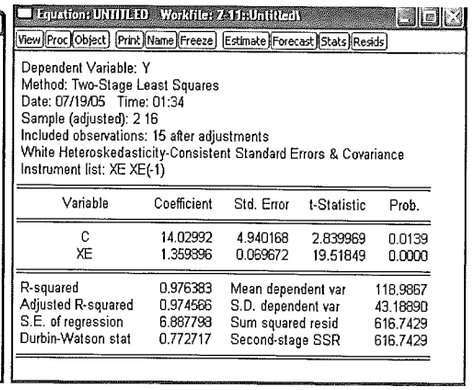


Figura 4-149

Se observa que el ajuste por mínimos cuadrados en dos etapas puede presentar autocorrelación y los parámetros estimados son casi iguales que los estimados anteriormente de modo directo por variables instrumentales en la Figura 4-145. Por tanto, es lógico quedarse con la estimación de la Figura 4-145.

Para analizar la posible heteroscedasticidad en el modelo de variable instrumental de la Figura 4-145 utilizaremos el ***contraste de White***. Para ello, desde la pantalla *Equation* mediante *View* → *Residual Tests* → *White Heteroskedasticity* (*no cross terms*) (Figura 4-150) se obtiene la Figura 4-151 en la que se observa que los p-valores de la *F* y de los términos cruzados son mayores que 0,05, lo que nos lleva a rechazar formalmente la presencia de heteroscedasticidad al 95%.

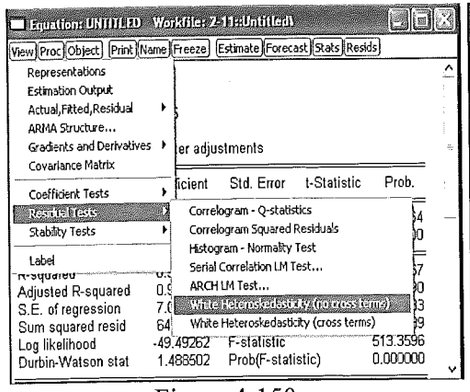


Figura 4-150

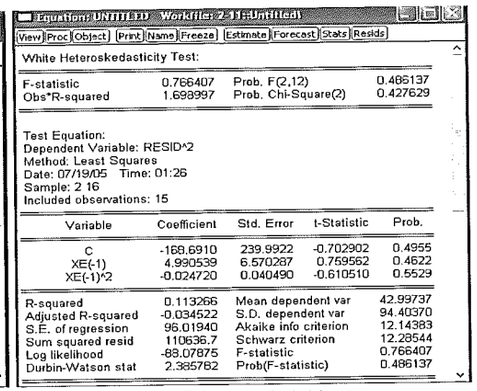


Figura 4-151

Ejercicio 4-3. El archivo 4-3.sav contiene datos sobre un modelo de demanda de productos agrícolas. El modelo involucra las variables DEMANDA, PRECIO, INGRESOS y PRECIPIT que indican el nivel de demanda y el precio del producto, así como los ingresos del consumidor y el nivel de precipitaciones respectivamente. Se trata de ajustar a los datos un modelo adecuado que explique la demanda del producto en función de su precio y los ingresos del consumidor.

Al tratarse de un modelo de demanda pueden existir problemas de regresores estocásticos, es decir, que los errores de la variable dependiente estén correlacionados con la variable o variables independientes. Por tanto ésta será la primera hipótesis que habrá que comprobar. Para ello, comenzamos realizando el ajuste del modelo (con el objeto de calcular los residuos), rellenado la pantalla de entrada del procedimiento *Regresión lineal* (*Analizar* → *Regresión* → *Lineal*) como se indica en la Figura 4-152 y su pantalla *Guardar* como se indica en la Figura 4-153, para guardar los residuos estudentizados como una nueva variable.

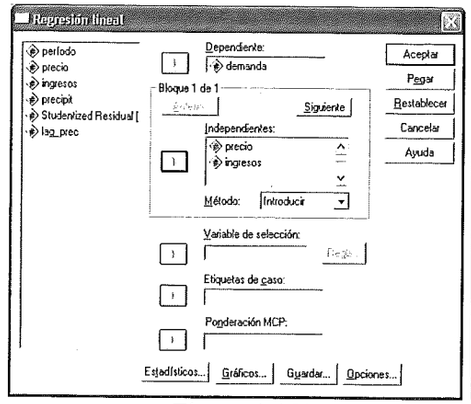


Figura 4-152

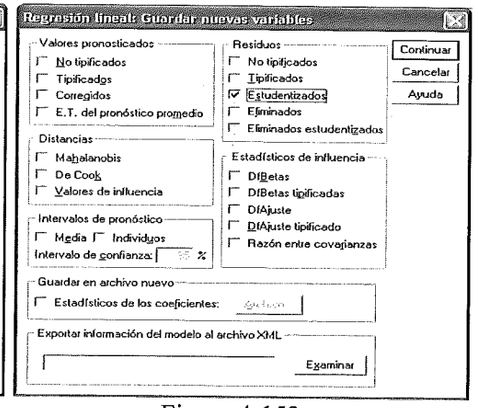


Figura 4-153

Al pulsar *Aceptar* la variable *sre_1* que contiene los residuos se incorpora al editor de SPSS (Figura 4-154). A continuación, se halla la matriz de correlaciones de las variables independientes y los residuos rellenando la pantalla de entrada del procedimiento *Correlaciones bivariadas* (*Analizar* → *Correlaciones* → *Bivariadas*) tal y como se indica en la Figura 4-155. Al pulsar *Aceptar* se obtiene la matriz de correlaciones de la Figura 4-156, que indica la fuerte correlación de los residuos con las variables independientes, lo que viola uno de los supuestos esenciales del modelo lineal mediante la existencia de regresores estocásticos.

demanda	precio	ingresos	precipit	sre_1
11	20	8,1	42	-,68065
16	18	8,4	58	,61312
11	22	8,5	35	-1,29058
14	21	8,5	46	,08556
13	27	8,8	41	,86194
17	26	9,0	56	1,93018
14	25	8,9	48	,16890
15	27	9,4	50	-,29503
12	30	9,5	39	-1,41740
18	28	9,9	52	,05592

Figura 4-154

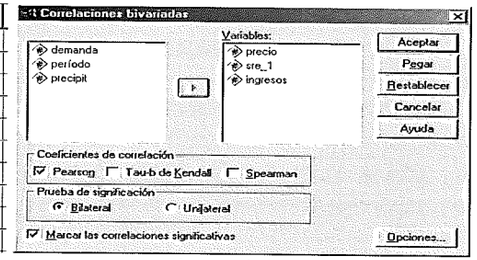


Figura 4-155

Correlaciones

		precio	ingresos	Studentized Residual
precio	Correlación de Pearson	1	,868**	-,034
	Sig. (bilateral)		,001	,926
	N	10	10	10
ingresos	Correlación de Pearson	,868**	1	-,023
	Sig. (bilateral)	,001		,950
	N	10	10	10
Studentized Residual	Correlación de Pearson	-,034	-,023	1
	Sig. (bilateral)	,926	,950	
	N	10	10	10

**. La correlación es significativa al nivel 0,01 (bilateral).

Figura 4-156

Ante la presencia de errores estocásticos, para ajustar nuestro modelo utilizamos el procedimiento *Regresión por Mínimos cuadrados en dos fases* (*Analizar* → *Regresión* → *Mínimos cuadrados en dos fases*) rellenando su pantalla de entrada como se indica en la Figura 4-157.

Utilizamos como instrumentos la variable predictora *ingresos* ya que es posible que los niveles de ingreso sean usados para predecir niveles de precios, la variable precios desfasada *lag_prec* (generada con *Transformar* → *Calcular* mediante LAG(precios), que suele ser un buen predictor del valor corriente de la demanda y la variable *precipit* que sabemos que influye sobre la demanda de alcachofas (Figura 4-158).

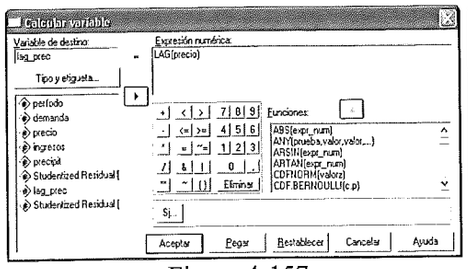


Figura 4-157

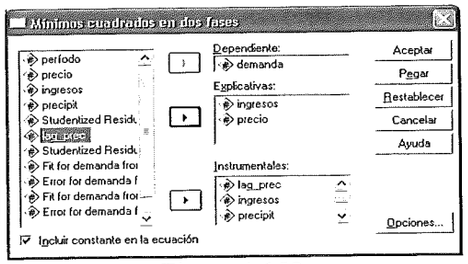


Figura 4-158

Al pulsar *Aceptar* se obtiene el ajuste de la Figura 4-159.

Resumen del modelo

Ecuación 1	R múltiple	,681
	R cuadrado	,464
	R cuadrado corregida	,285
	Error típico de la estimación	2,442

ANOVA

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
Ecuación 1	Regresión	30,977	2	15,489	2,598	,154
	Residuo	35,777	6	5,963		
	Total	66,754	8			

Coeficientes

		Coeficientes no estandarizados		Beta	t	Sig.
		B	Error típico			
Ecuación 1	(Constante)	-40,017	24,452		-1,637	,153
	ingresos	9,561	4,199	2,147	2,277	,063
	precio	-1,265	,617	-2,105	-2,051	,086

Figura 4-159

La significatividad de las variables supera el 90% y la significatividad de la constante, así como la significatividad conjunta, se aproxima al 85%. El modelo ajustado tiene la siguiente expresión:

$$Demanda = -40{,}016582 + 9{,}561323\ ingresos - 1{,}265008\ precio$$

Lógicamente la demanda es directamente proporcional a los ingresos (signo positivo del coeficiente estimado) e inversamente proporcional al precio (signo negativo del coeficiente estimado), siendo la influencia de los ingresos sobre la demanda del producto bastante superior a la del precio de dicho producto.

Ejercicio 4-4. Se trata de explicar la cantidad demandada Q1 del bien 1 en función de su precio P1, del nivel de renta disponible YD y de los precios P2 y P3 de dos bienes alternativos mediante el modelo $Q1 = b1*P1^{b2}*P2^{b3}*P3^{b4}*YD^{b5}$***. Se dispone de los datos almacenados en el fichero en formato eviews 4-4.wk1.***

Ajustar de forma razonable el modelo analizando los posibles problemas.

El primer paso para trabajar con este modelo es linealizarlo aplicando logaritmos, con lo que se transformará en el modelo:

$$Log(Q1) = Log(b_1) + b_2*Log(P1) + b_3*Log(P2) + b_4*Log(P3) + b5*Log(YD) + u$$

Realizaremos en primer lugar el ajuste por mínimos cuadrados de nuestro modelo. Esta tarea puede automatizarse con el software econométrico Eviews introduciendo los datos en un fichero de trabajo (4-4.*wk*1), cargándolo en el programa usando *File* → *Open* → *Eviews Workfile...* y eligiendo el fichero en la ventana *Open*. Al pulsar *Abrir* se carga en memoria el fichero. Si queremos ver el contenido del conjunto de datos bastará con seleccionar todas las variables del modo habitual en Windows, hacer clic con el botón derecho del ratón sobre la selección y elegir *Open* → *As Group*. Se ve así el contenido de todas las variables del modelo como un grupo (Figura 4-160).

Para ***realizar el ajuste MCO***, se elige *Quick* → *Estimate Equation* (Figura 4-160), se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *Least Squares* en el campo *Method* para ajustar por mínimos cuadrados (Figura 4-161) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-162.

Inicialmente el ajuste parece correcto, pues hay significatividad al menos al 93% (p-valores de la T menores que 0,07) de todos los parámetros y también hay fuerte significatividad global (p-valor de la F casi nulo). Por otra parte no hay autocorrelación, ya que el estadístico de Durbin-Watson se acerca mucho a 2. Quizás el R^2 sea algo bajo, pero no drásticamente (65,4%).

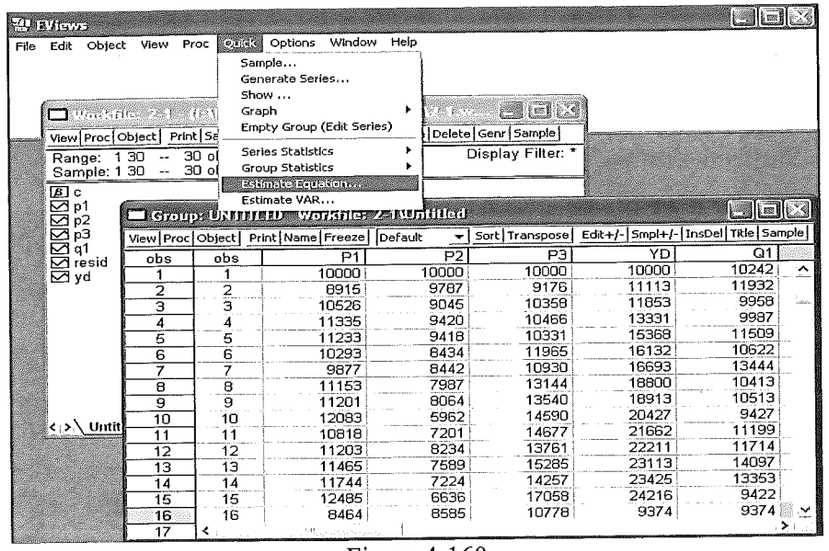


Figura 4-160

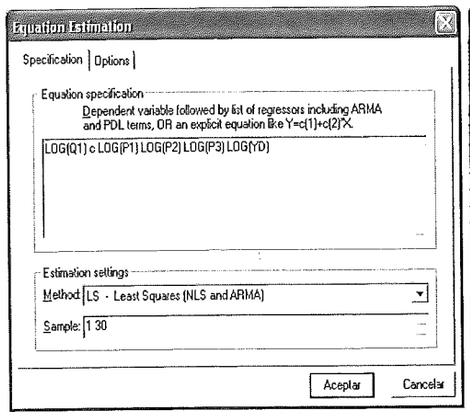


Figura 4-161

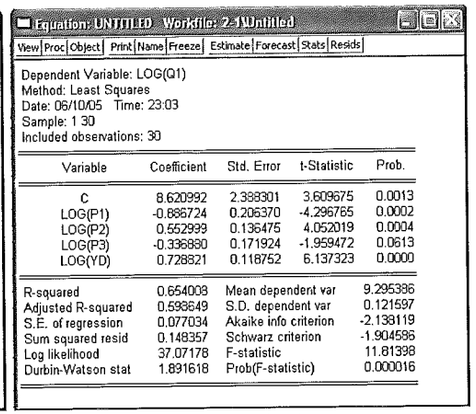


Figura 4-162

Pueden representarse los residuos contra los valores predichos para analizar la posible heteroscedasticidad. Para ello se calculan los valores predichos eligiendo *Proc* → *Forecast* en la pantalla *Equation* (Figura 4-163) y rellenando la pantalla *Forecast* como se indica en la Figura 4-164 (El vector *LOGq1f* contiene las predicciones de la variable dependiente). A continuación se elige *Quick* → *Graph* → *Scatter* (Figura 4-165) y se rellena la pantalla *Series List* como se indica en la Figura 4-166. Al pulsar *OK* se obtiene la Figura 4-167 que también presenta una estructura poco aleatoria de sus puntos, lo que indica presencia de ***problemas de heteroscedasticidad.***

Para detectar qué variables son las responsables de la posible heteroscedasticidad realizamos los gráficos de residuos contra las cuatro variables explicativas. Para ello elegimos *Quick* → *Graph* → *Scatter* (Figura 4-165) y rellenamos la pantalla *Series List* sucesivamente como se indica en las Figuras 4-168, 4-170, 4-172 y 4-174. Al pulsar *OK* se obtienen las Figuras 4-169, 4-171, 4-173 y 4-175. Se observa que la menos aleatoria es la relativa a *YD*, ya que aumenta la dispersión del error al ir de izquierda a derecha y presenta un ajuste bueno a una recta paralela al eje *X*; por tanto no tiene estructura aleatoria.

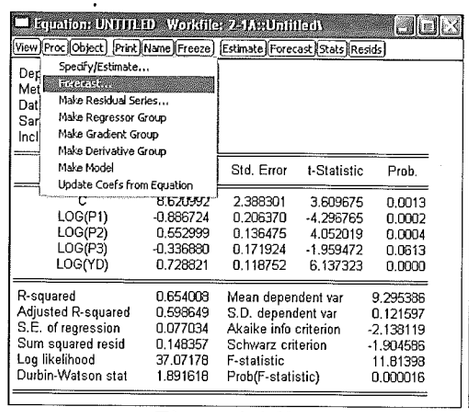


Figura 4-163

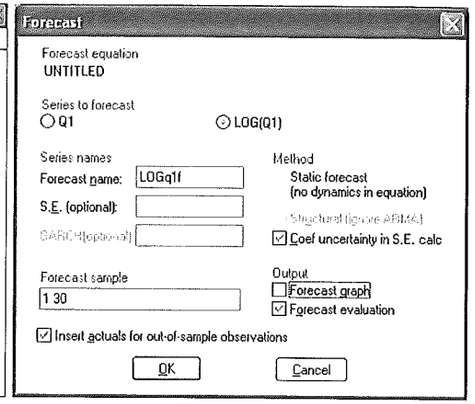


Figura 4-164

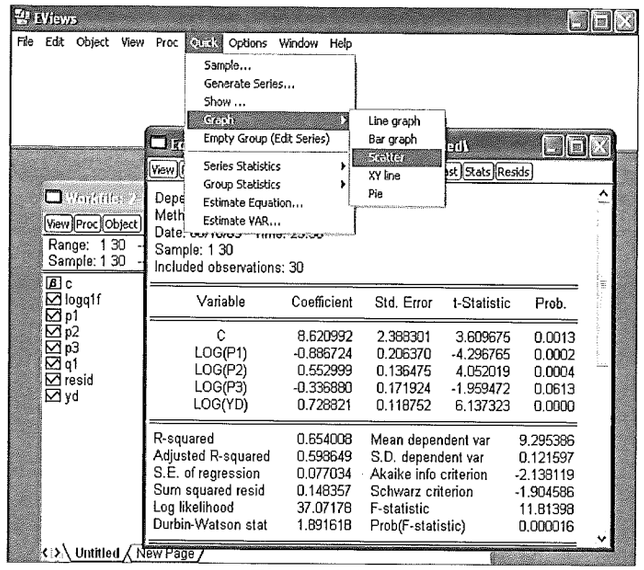


Figura 4-165

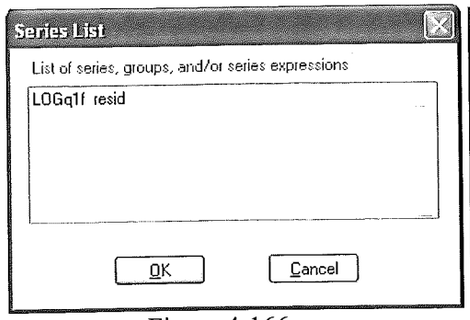


Figura 4-166

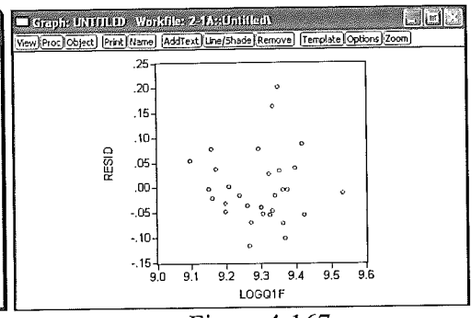


Figura 4-167

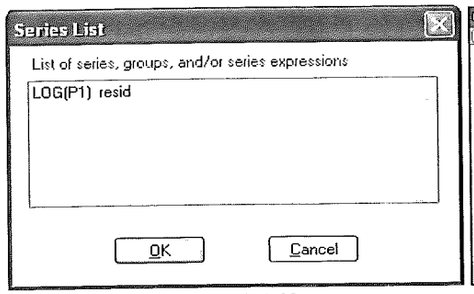


Figura 4-168

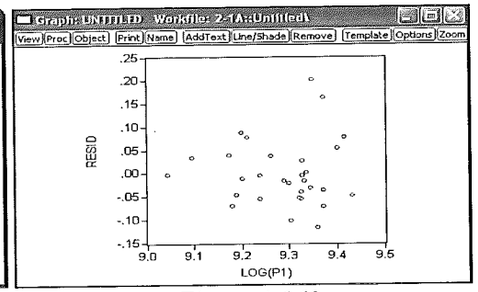


Figura 4-169

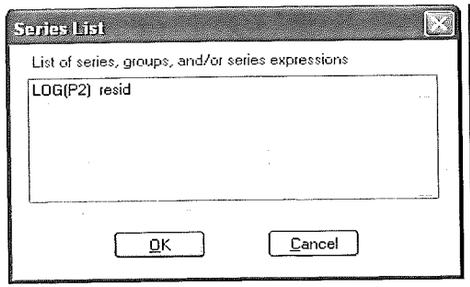


Figura 4-170

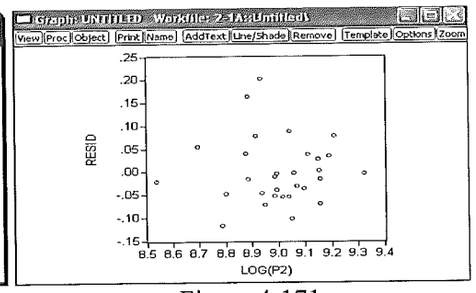


Figura 4-171

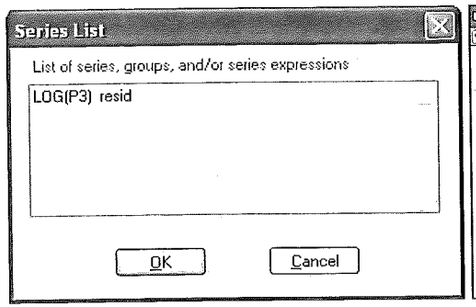


Figura 4-172

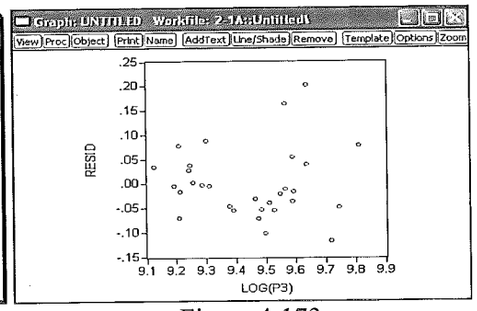


Figura 4-173

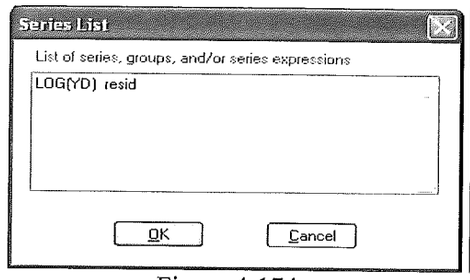


Figura 4-174

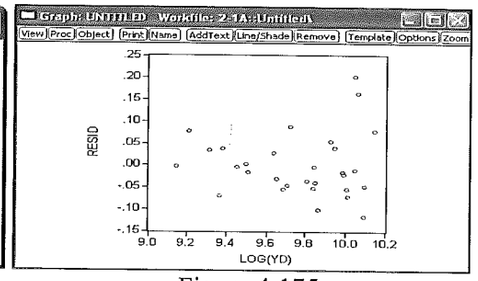


Figura 4-175

La variable *YD* es la candidata a provocar los problemas de heteroscedasticidad, pero para estar más seguros utilizaremos un contraste formal, por ejemplo el ***contraste de Glesjer***. Para ello realizamos una regresión múltiple de *Abs(RESID)* contra *Log(YD)* rellenando la pantalla *Equation Estimation* obtenida con *Quick* → *Estimate equation*) como se indica en la Figura 4-176. Al pulsar *Aceptar* se obtienen los resultados de la Figura 4-177.

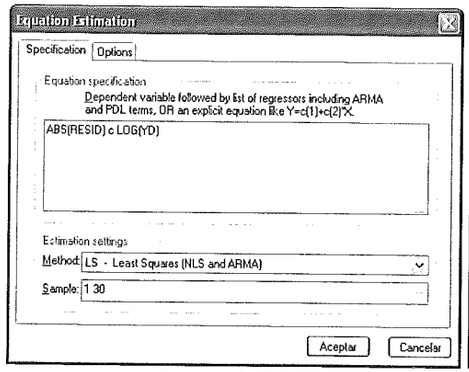


Figura 4-176

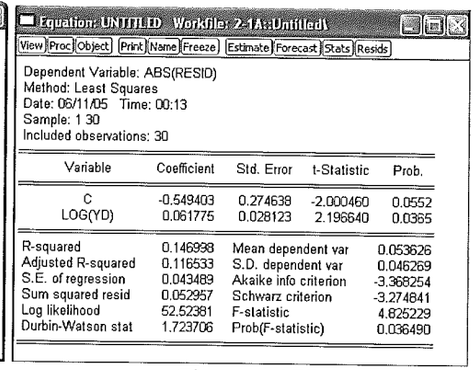


Figura 4-177

Todos los coeficientes resultan significativos al 94% (p-valores menores que 0,06) y el modelo global tambiñén es significativo (p-valor de la *F* menor que 0,5). Por tanto tenemos que rechazar la nulidad de cualquier coeficiente incluida la constante. Como, según el contraste de Glesjer, la hipótesis de homoscedasticidad es equivalente a la hipótesis de pendiente nula en este modelo, hay heteroscedasticidad según *YD*.

Ahora resta la tarea de estimar el modelo inicial con heteroscedasticidad. Para ello observamos en la figura de *YD* contra los residuos, que la dispersión residual es creciente con *YD*, lo que nos permite suponer como modelo para varianza residual $\sigma_i^2 = \sigma^2(log(YD))^2$, lo que nos permitirá estimar el modelo con heteroscedasticidad mediante ***mínimos cuadrados ponderados*** usando como variable de ponderación $1/log(YD)$.

Para ello rellenamos la solapa *Specification* de la pantalla *Equation Estimation* (obtenida con *Quick* → *Estimate equation*) como se indica en la Figura 4-178. La solapa *Options* se rellena como se indica en la Figura 4-179 marcando la opción *Weighted LS/TLS* y situando en el campo *Weigth* la variable *1/log(YD)*. Al pulsar *Aceptar* se obtienen los resultados de la Figura 4-180.

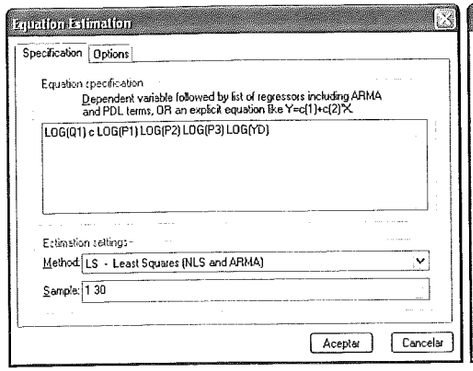


Figura 4-178

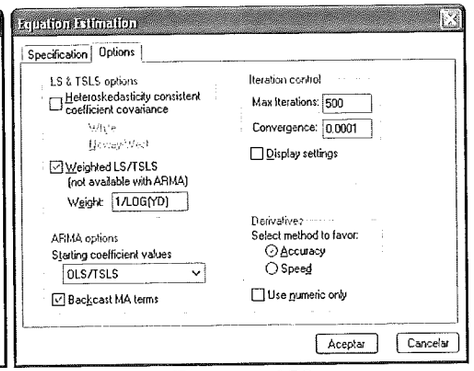


Figura 4-179

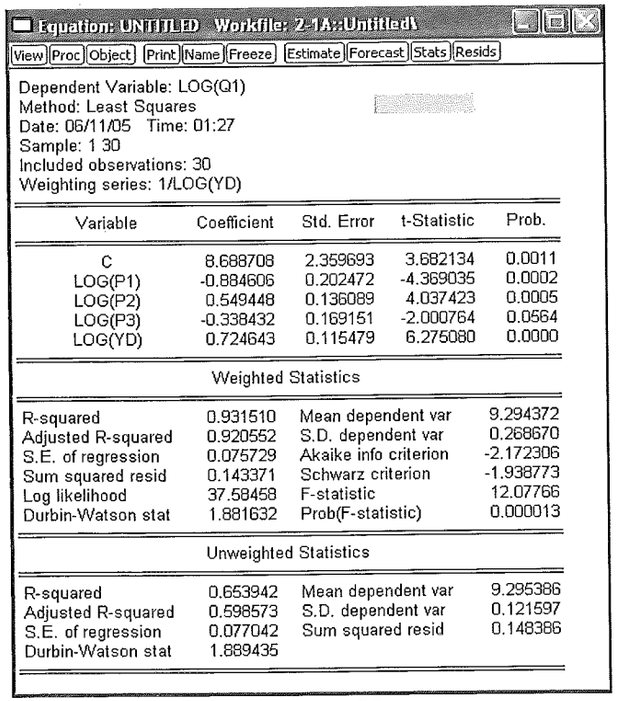


Figura 4-180

Se observa que el ajuste ponderado es bueno y que los parámetros estimados difieren levemente de los obtenidos antes de corregir la heteroscedasticidad. Ha mejorado R^2 y las significatividades de los parámetros estimados también han mejorado. El modelo estimado es el siguiente:

$$Log(Q1) = 8{,}6887 - 0{,}8846Log(P1) + 0{,}5494Log(P2) - 0{,}3384Log(P3) + 0{,}7246Log(YD) + u$$

Ejercicio 4-5. Mediante un estudio con datos de las Comunidades Autónomas españolas referentes a 1991, se trata de estimar la relación entre el nivel de empleo (medido por las medias anuales de ocupados en cada comunidad en miles de personas) y el PIB (medido por el valor añadido bruto al coste de los factores en las comunidades). Los datos se encuentran en el archivo 4-5.wf1. Estudiar los problemas posibles de heteroscedasticidad en el modelo y resolverlos de la forma más adecuada.

Realizaremos en primer lugar el ajuste por mínimos cuadrados de nuestro modelo. Esta tarea puede automatizarse con el software econométrico Eviews utilizando los datos del fichero de trabajo (4-5.*wf1*), cargándolo en el programa usando *File* → *Open* → *Eviews Workfile...* y eligiendo el fichero en la ventana *Open*. Al pulsar *Abrir* se carga en memoria. Si queremos ver el contenido del conjunto de datos bastará con seleccionar todas las variables del modo habitual en Windows, hacer clic con el botón derecho del ratón sobre la selección y elegir *Open* → *As Group*. Se ve así el contenido de todas las variables del modelo como un grupo (Figura 4-181).

Para ***realizar el ajuste MCO***, se elige *Quick* → *Estimate Equation* (Figura 4-22), se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *Least Squares* en el campo *Method* para ajustar por mínimos cuadrados (Figura 4-182) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-183.

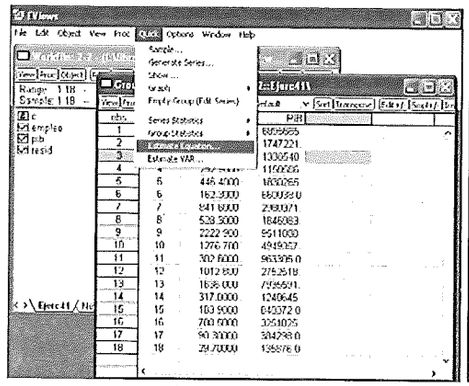


Figura 4-181

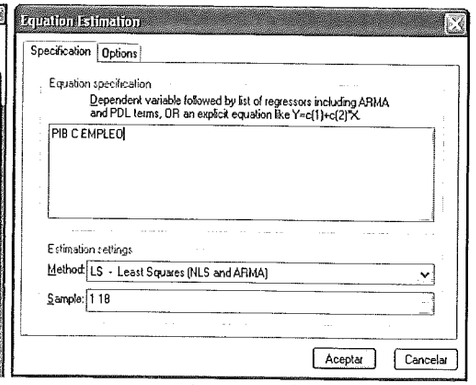


Figura 4-182

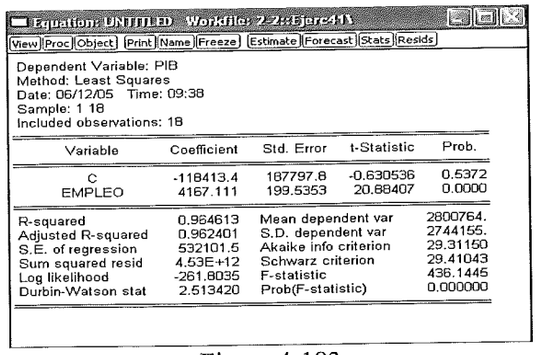


Figura 4-183

Inicialmente el ajuste presenta no significatividad acusada de la constante y una significatividad muy alta tanto conjunta como de la variable empleo. Por otra parte parece que no hay autocorrelación, ya que el estadístico de Durbin-Watson no se aleja demasiado de 2. El R^2 también es muy alto (96%).

Probaremos ahora si hay ***problemas de heteroscedasticidad***. Para ello aplicamos el ***contraste de Goldfed-Quandt***, que requiere la ordenación previa de forma ascendente de las variables del modelo según los valores de la variable que provoca la heteroscedasticidad (en nuestro caso *empleo*). Para ello utilizamos *Proc → Sort Current page* (Figura 4-184) indicando *empleo* como variable de ordenación en la pantalla *Sort Workfile Series* (Figura 4-185).

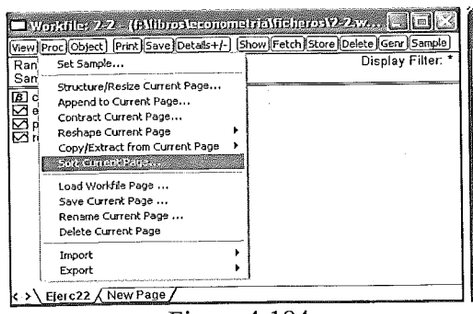


Figura 4-184

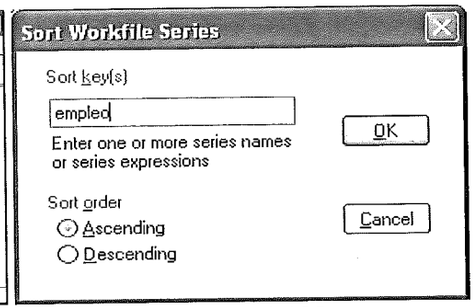


Figura 4-185

Ahora se ajusta el modelo con las primeras $(n\text{-}c)/2$ observaciones de la muestra ordenada con $c = n/3 = 18/3 = 6$. Se tiene que $(n\text{-}c)/2 = (18\text{-}6)/2 = 6$, con lo que la primera submuestra constará de las primeras 6 observaciones. El ajuste del modelo con estas 6 observaciones se realiza con Eviews mediante *Quick → Estimate Equation* y rellenando el campo *Equation Specification* de la solapa *Specification* como se indica en la Figura 4-186. Se elige *Least Squares* en el campo *Method* para ajustar por mínimos cuadrados, en el campo Sample se sitúa el rango muestral 1 6 y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-187.

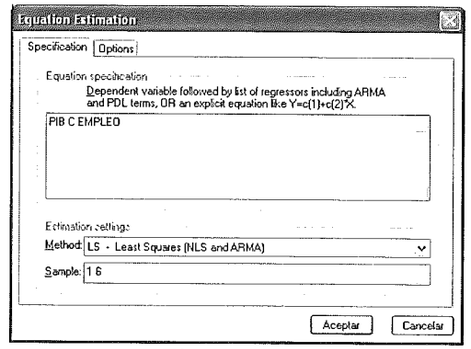


Figura 4-186

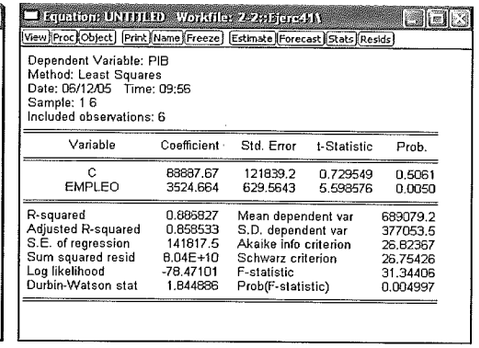


Figura 4-187

A continuación se ajusta el modelo con las últimas $(n\text{-}c)/2$ observaciones de la muestra ordenada, es decir con la últimas $(n\text{-}c)/2 = (18\text{-}6)/2 = 6$ observaciones, es decir de la 13 a la 18. El ajuste del modelo con estas 6 observaciones se realiza con Eviews mediante *Quick → Estimate Equation* y rellenando el campo *Equation Specification* de la solapa *Specification* como se indica en la Figura 4-188. Se elige *Least Squares* en el campo *Method* para ajustar por mínimos cuadrados, en el campo Sample se sitúa el rango muestral 13 18 y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-189.

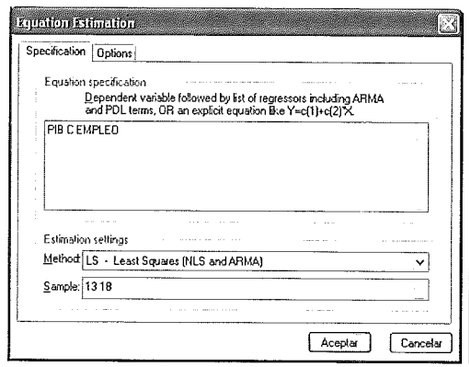


Figura 4-188

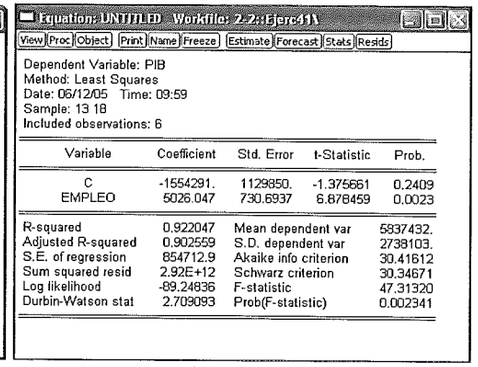


Figura 4-189

Para obtener el valor del estadístico del contraste de Goldfeld-Quandt, se dividen los valores de las sumas de cuadrados residuales de las dos regresiones anteriores =2.92E+12/8.04E+10 usando la línea de comandos de Eviews (Figura 4-190). El resultado (36,3184) se obtiene en la línea de estado de la parte inferior de la pantalla (Figura 4-190). Ahora hay que calcular el valor de la función de distribución de una F de Fisher Snedokor $((n\text{-}c)/4\text{-}k,\ (n\text{-}c)/4\text{-}k)$ en el punto 36,3184. Se escribe en la línea de comandos =@FDIST(36.3184,4,4) cuyo valor es 0,002115 (Figura 4-191). Como 0,002115<0,05 se rechaza la hipótesis nula de homoscedasticidad. Hemos detectado así formalmente la presencia de heteroscedasticidad.

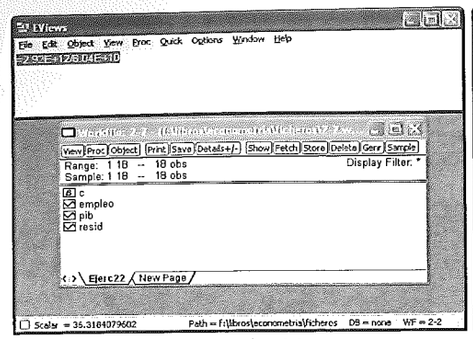


Figura 4-190

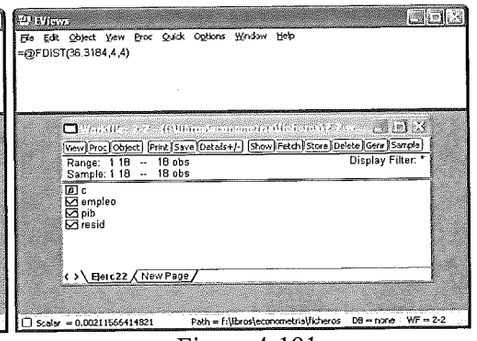


Figura 4-191

Para ajustar correctamente el modelo en presencia de heteroscedasticidad utilizaremos el ***método de estimación de White***. Para ello rellenamos la solapa *Specification* de la pantalla *Equation Estimation* (obtenida con *Quick* → *Estimate equation*) como se indica en la Figura 4-192. La solapa *Options* se rellena como se indica en la Figura 4-193 eligiendo la opción *White* en el campo *Heteroskedasticity consistent coeficient covariance*. Al pulsar *Aceptar* se obtienen los resultados de la Figura 4-194. Se observa que la significatividad de la constante ha mejorado y que el resto de los indicadores de ajuste del modelo son buenos: R^2 muy alto, estadístico de Durbin Watson cercano a 2 y criterios de información de Akaike y Swartz con valores bajos.

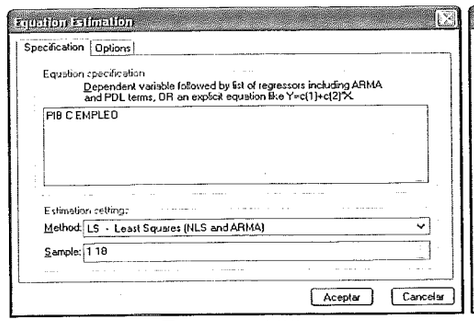


Figura 4-192

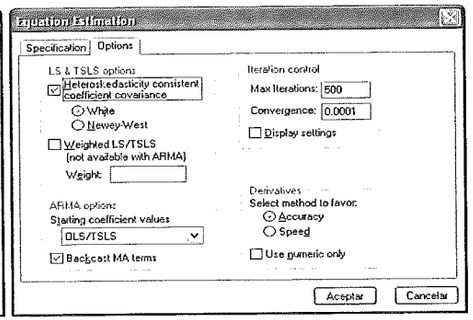


Figura 4-193

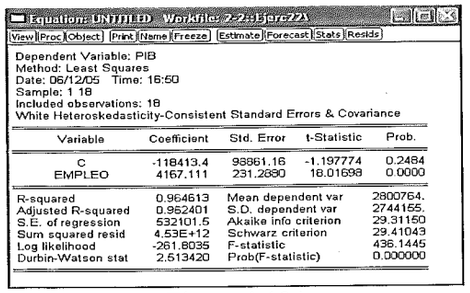


Figura 4-194

Ejercicio 4-6. El conjunto de datos 4-6.wf1 contiene información agregada relativa al consumo final de las familias residentes (Y) y la renta bruta disponible de las mismas (X) por Comunidades Autónomas para el año 1991. Con vistas a una posible construcción de grandes superficies, un grupo empresarial quiere conocer como se explica el consumo final de las familias en función de su renta bruta disponible. Analizar el modelo salvando los posibles problemas de heteroscrdasticidad.

Se trata de ajustar de forma razonable el modelo siguiente:

$$Y = \beta_0 + \beta_1 X + u$$

Realizaremos en primer lugar el ajuste por mínimos cuadrados ordinarios de nuestro modelo cargando el fichero 4-6.*wf1* mediante *File* → *Open* → *Eviews Workfile...* y eligiendo el fichero en la ventana *Open*. Al pulsar *Abrir* se carga en memoria el fichero. Para ***realizar el ajuste MCO***, se elige *Quick* → *Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *Least Squares* en el campo *Method* para ajustar por mínimos cuadrados (Figura 4-195) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-196.

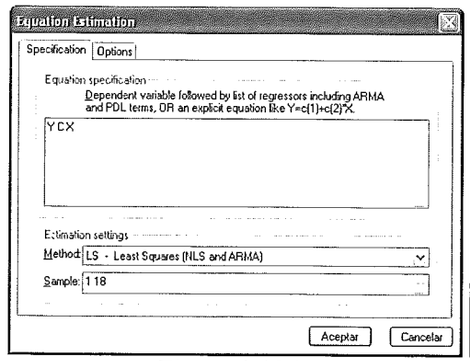


Figura 4-195

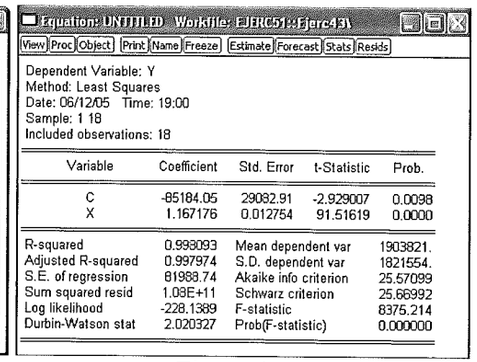


Figura 4-196

Inicialmente el ajuste parece correcto, pues hay significatividad al menos al 99% (p-valores de la T menores que 0,01) de todos los parámetros y también hay fuerte significatividad global (p-valor de la F casi nulo). Por otra parte no hay autocorrelación, ya que el estadístico de Durbin-Watson se acerca mucho a 2. El R^2 es muy alto (superior al 99%).

Pero dado que estamos trabajando con datos de corte transversal siempre existe el peligro de la heteroscedasticidad. Para detectar inicialmente la posible heteroscedasticidad realizamos el gráfico de residuos contra la única variable explicativa, que lógicamente sería la responsable de la ***posible heteroscedasticidad***.

Par ello elegimos *Quick* → *Graph* → *Scatter* (Figura 4-197) y rellenamos la pantalla *Series List* como se indica en la Figura 4-198. Al pulsar *OK* se obtiene la Figura 4-199. Se observa que la distribución de los puntos del gráfico no es aleatoria. Ello nos lleva a ensayar contrastes formales de heteroscedasticidad.

Utilizaremos inicialmente el ***contraste de White*** consistente en efectuar una regresión de los cuadrados de los residuos MCO sobre todas las variables independientes del modelo, sus cuadrados y sus productos cruzados de dos en dos. La homoscedasticidad se acepta al 95% si los p-valores de la F y de los términos cruzados son mayores que 0,05.

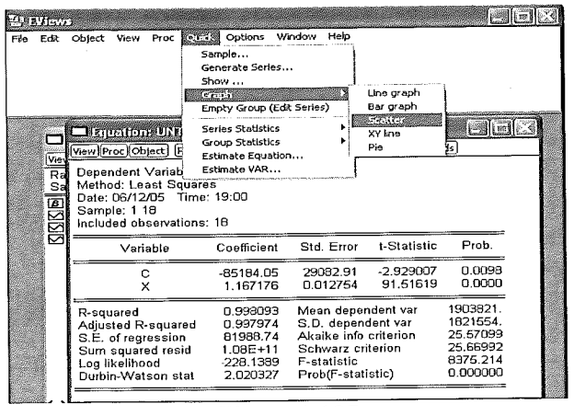


Figura 4-197

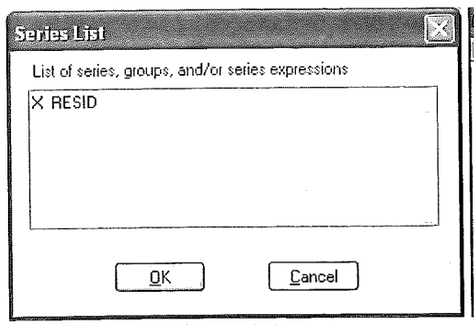


Figura 4-198

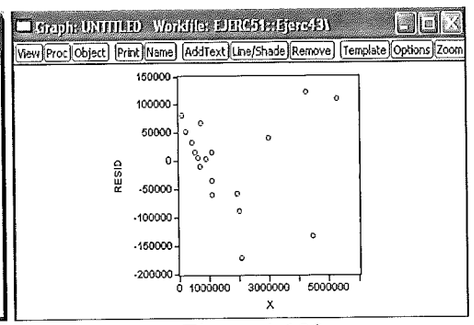


Figura 4-199

Eviews permite realizar el contraste de White de heteroscedasticidad desde la pantalla *Equation* mediante *View* → *Residual Tests* → *White Hetereroskedasticity* (*cross terms*) (Figura 4-200). Se obtiene la Figura 4-201 en la que se observa que los p-valores de la F y de los términos cruzados son menores que 0,05, lo que nos lleva a aceptar formalmente la presencia de heteroscedasticidad al 95%.

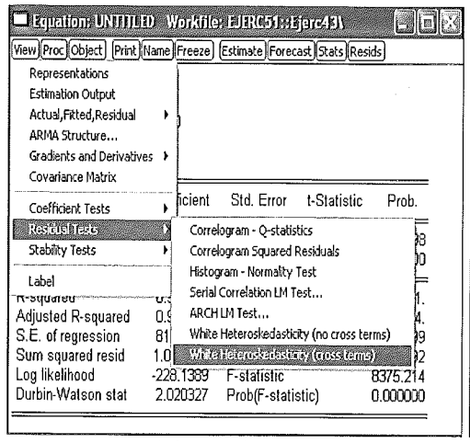


Figura 4-200

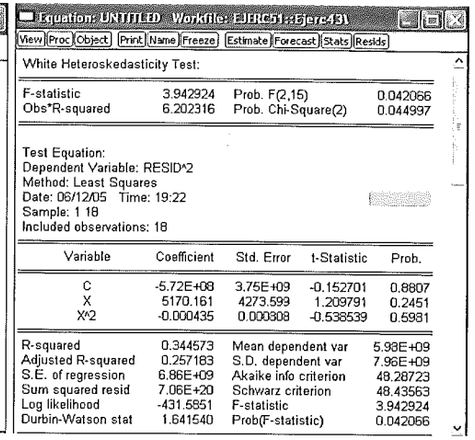


Figura 4-201

Adicionalmente podemos utilizar el ***contraste de Breusch-Pagan***, que se realiza obteniendo los residuos del ajuste del modelo y considerando los cocientes g_i entre los cuadrados de cada residuo del modelo original (RESID^2) y la estimación MV de la varianza del error (suma de los cuadrados de los residuos/n = 1,08E+11/18 =6000000000). Al realizar la regresión entre las variables g_i y las variables culpables Z_i (en nuestro caso sólo la variable X) más la constante, la cantidad Q = SCE/2 se distribuye según una Chi-cuadrado con p-1 = 1 grados de libertad bajo la hipótesis nula de homoscedasticidad. SCE es la varianza explicada de la regresión anterior (suma de cuadrados explicados). Para realizar esta regresión con Eviews, se elige *Quick* → *Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *Least Squares* en el campo *Method* para ajustar por mínimos cuadrados (Figura 4-202) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-203.

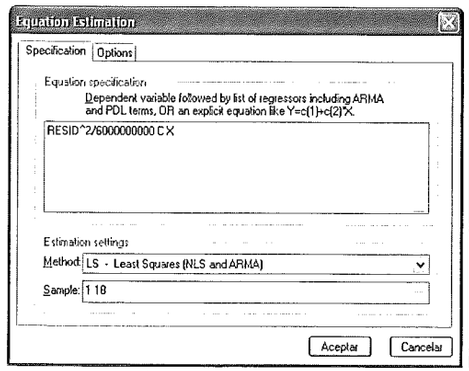


Figura 4-202

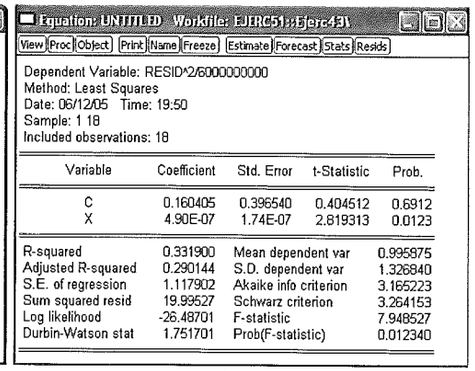


Figura 4-203

La varianza explicada de la regresión anterior se calcula como $SCE = SCT - SCR = SCR/(1-R^2) - SCR = 19{,}99527/(1-0{,}3319) - 19{,}99527 = 9{,}933288599$ (Figura 4-204). *SCE* es la suma de cuadrados explicados, *SCT* es la suma de cuadrados total y SCR es la suma de cuadrados residual. El estadístico de Breusch-Pagan es $SCE/2 = 4{,}96666442995$ y el p-valor será el valor de la función de distribución de una chi-cuadrado con un grado de libertad en el punto 9,933288599, que se calcula con Eviews introduciendo en la línea de comandos =@CHISQ(4. 96666442995,1) cuyo valor es 0,02584 (Figura 4-205). Como el p-valor es menor que 0,05 se acepta la hipótesis de heteroscedasticidad.

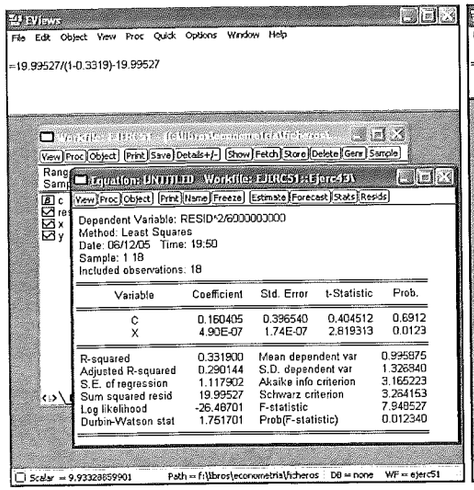


Figura 4-204

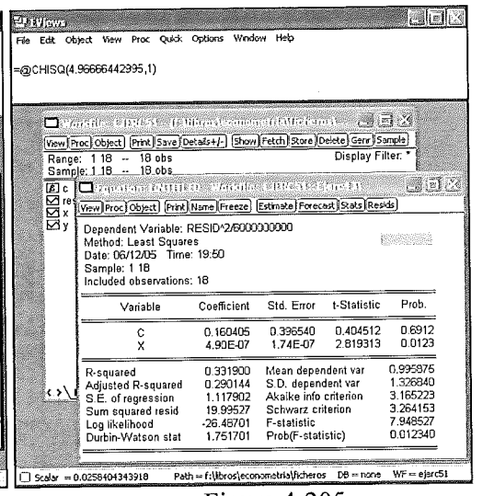


Figura 4-205

Una vez detectada la presencia de heteroscedasticidad en el modelo es necesario corregirla en la estimación del mismo. Para ajustar correctamente el modelo en presencia de heteroscedasticidad utilizaremos el ***método de estimación de White***. Para ello rellenamos la solapa *Specification* de la pantalla *Equation Estimation* (obtenida con *Quick* → *Estimate equation*) como se indica en la Figura 4-206. La solapa *Options* se rellena como se indica en la Figura 4-207 eligiendo la opción *White* en el campo *Heteroskedasticity consistent coefficient covariance*. Al pulsar *Aceptar* se obtienen los resultados de la Figura 4-208. Se observa que la significatividad de la constante ha mejorado y que el resto de los indicadores de ajuste del modelo son buenos: R^2 muy alto, estadístico de Durbin Watson cercano a 2 y criterios de información de Akaike y Swartz con valores bajos.

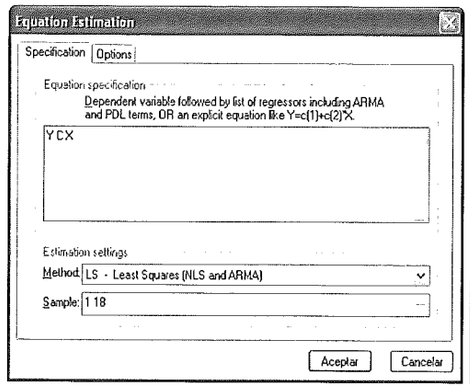


Figura 4-206

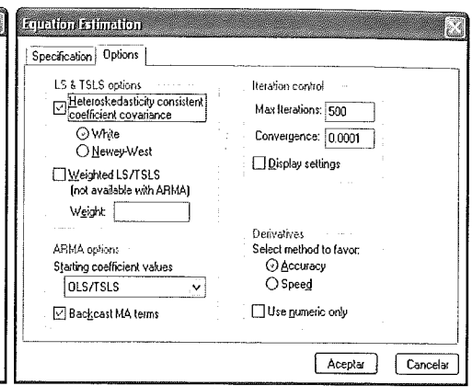


Figura 4-207

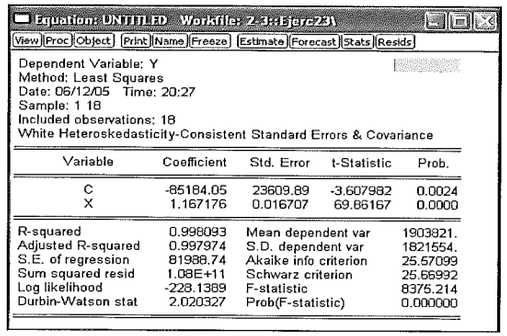


Figura 4-208

Finalmente el modelo puede ajustarse como sigue:

$$Y = -85184{,}05 + 1{,}167176\,X + u$$

Al interpretar los resultados del ajuste podemos decir que el aumento de una unidad en la renta disponible de las familias residentes en España en 1991 produce un aumento de 1,167176 unidades en el consumo final de las citadas familias. Ello podría llevar al grupo empresarial a tomar la decisión de la construcción de grandes superficies.

Ejercicio 4-7. El archivo 4-7.wf1 contiene información de la economía española correspondiente al período 1964-1980 sobre tres variables macroeconómicas: importaciones energéticas reales (IMPEN), producto interior bruto a precios de mercado (PIB, (ambas en miles de millones de pesetas constantes de 1970), y el precio relativo de las importaciones energética (PREN)

Estimar el modelo de regresión que pretende explicar el volumen de importaciones energéticas en función de las variaciones del PIB, y el precio relativo de las importaciones. Criticar el modelo e introducir alguna variable ficticia adecuada con criterio económico que ayude a obtener un buen ajuste.

Si ajustamos el modelo por MCO rellenando la pantalla *Equation Estimation* de Eviews como se indica en la Figura 4-209 se obtienen los resultados de la Figura 4-210. Se observa buena significatividad individual y conjunta de los parámetros y buen coeficiente de determinación, pero el valor del estadístico de Durbin Watson inicialmente lejano de 2 puede indicar presencia de autocorrelación.

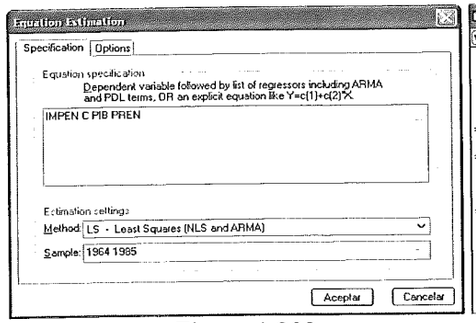


Figura 4-209

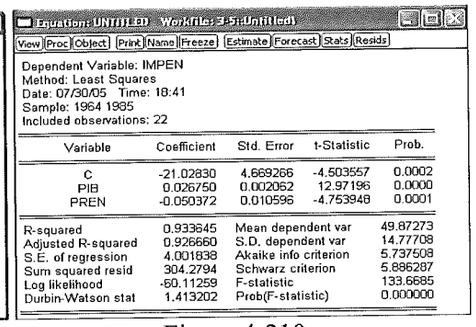


Figura 4-210

En la tabla del estadístico de Durbin-Watson para $T = 22$, $k = 2$ y $\alpha = 0{,}05$ se tiene que $d_L = 1{,}147$ y $d_U = 1{,}541$. Como $d_L = 1{,}147 < d = 1{,}41 < d_U = 1{,}541$ estamos en la zona de indeterminación de la autocorrelación.

Una alternativa para clarificar el problema de la autocorrelación podría ser la ***introducción de una variable ficticia***. El posible criterio económico a utilizar podría ser el impacto de la crisis energética de 1997 en la economía española. Dicha variable, de nombre FICTICIA, valdrá 0 entre 1964 y 1976, y valdrá 1 entre 1977 y 1985. Al añadir esta variable y ajustar el modelo por MCO rellenando la pantalla *Equation Estimation* de Eviews como se indica en la Figura 4-211, se obtienen los resultados de la Figura 4-212. Se observa que ha mejorado la significatividad individual y conjunta de los parámetros y los coeficientes de determinación, y el valor del estadístico de Durbin Watson se acerca más a 2. En la tabla del estadístico de Durbin-Watson para $T = 22$, $k = 2$ y $\alpha = 0{,}05$ se tiene que $d_L = 1{,}05$ y $d_U = 1{,}66$. Como $d_U \leq$ d=1,662690 está más clara la ausencia de autocorrelación.

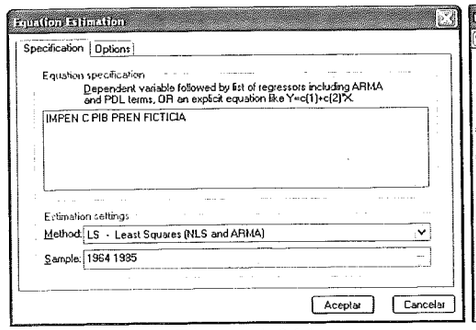


Figura 4-211

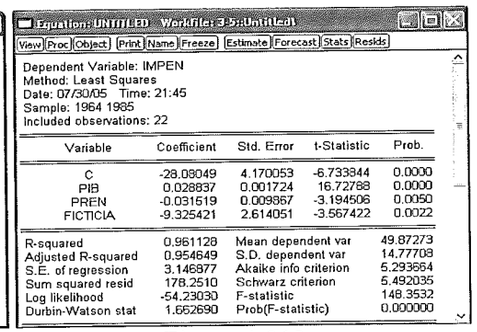


Figura 4-212

Los p-valores de la F y la Chi-cuadrado mayores que 0,05 del test de White (Figura 4-213 y 4-214) indican ausencia de heteroscedasticidad y el p-valor del test de Jarque-Vera mayor que 0,005 (Figura 4-215 y 4-216) indica normalidad de los residuos. El modelo ajustado será:

$$IMPEN = -28.080487 + 0.028836756*PIB - 0.03151911684*PREN - 9.325*FICTICIA + u$$

Los parámetros estimados difieren poco de las estimaciones obtenidas sin la variable ficticia ya que el problema de la autocorrelación no era muy acusado.

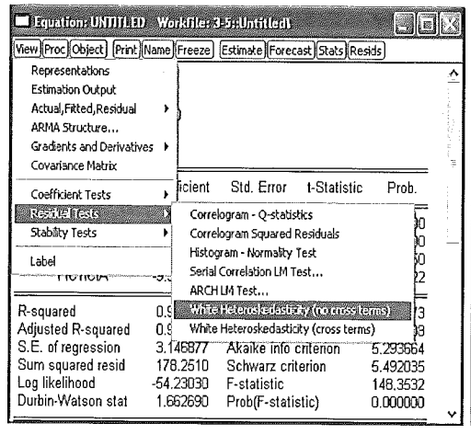


Figura 4-213

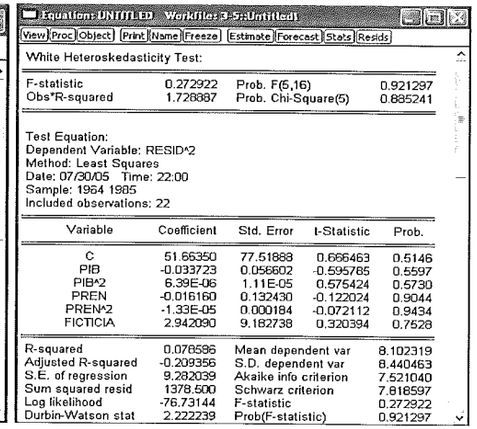


Figura 4-214

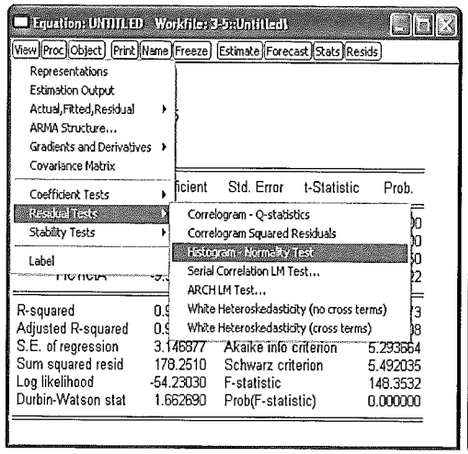


Figura 4-215

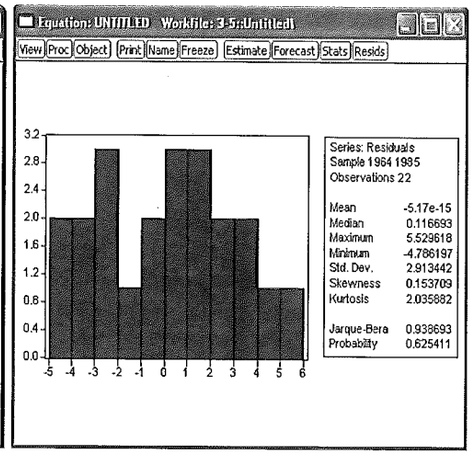


Figura 4-216

Ejercicio 4-8. Con el objetivo de estudiar la evolución fiscal en España se estima un modelo de regresión que intenta explicar las variaciones de los ingresos públicos a partir del PIB nominal. Se dispone del archivo 4-8.wf1 que contiene información del Banco de España referida al periodo 1955-1991.

Con las variables en logaritmos, estime un modelo de regresión simple entre los ingresos públicos (Ingresos) y el PIB nominal (PIB). Represente los residuos de la estimación con respecto a la variable endógena del modelo. ¿Justifican estos residuos la existencia de algún comportamiento irregular en los ingresos públicos no explicado por el PIB nominal y no incluido en el modelo estimado?

Incluir en la estimación una variable ficticia, de modo aditivo y multiplicativo, que toma el valor 1 desde el año 1978 en adelante. Realizar la misma estimación valorándola estadísticamente. Representar de nuevo el gráfico de residuos comparándolo con el del apartado anterior.

Si ajustamos el modelo por MCO rellenando la pantalla *Equation Estimation* de Eviews como se indica en la Figura 4-217 se obtienen los resultados de la Figura 4-218. Se observa significatividad individual y conjunta de los parámetros muy alta y muy buen coeficiente de determinación, pero el valor del estadístico de Durbin Watson inicialmente muy lejano de 2 puede indicar presencia de autocorrelación.

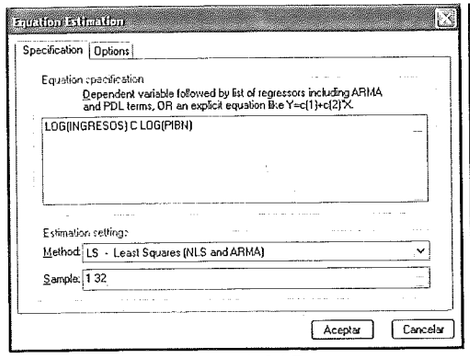


Figura 4-217

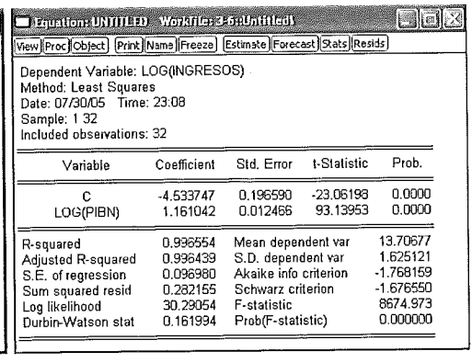


Figura 4-218

En la tabla del estadístico de Durbin-Watson para $T = 32$, $k = 1$ y $\alpha = 0{,}05$ se tiene que $d_L = 1{,}373$ y $d_U = 1{,}502$. Como $d = 0{,}161994 < d_L = 1{,}373$, existe autocorrelación positiva. Para representar los residuos contra la variable endógena del modelo utilizamos *Quick* → *Graph* → *Scatter*, rellenamos la pantalla *Series List* como se indica en la Figura 4-219 y al pulsar *OK* se obtiene el gráfico de la Figura 4-220, cuya estructura no aleatoria indica la no linealidad del modelo. Como además la gráfica es ajustable a una parábola, el modelo podrá contener términos de orden dos (cuadrados o productos de variables). Ello nos lleva a incluir la ***variable ficticia en el modelo, tanto aditivamente como multiplicativamente.***

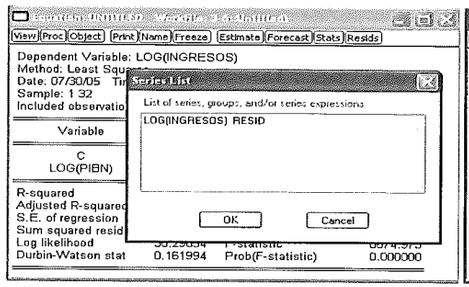


Figura 4-219

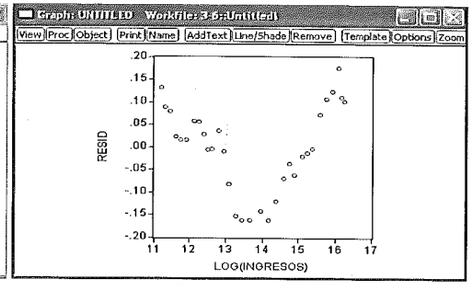


Figura 4-220

Al añadir la variable ficticia aditiva y multiplicativamente y ajustar el modelo por MCO rellenando la pantalla *Equation Estimation* de Eviews como se indica en la Figura 4-221, se obtienen los resultados de la Figura 4-222.

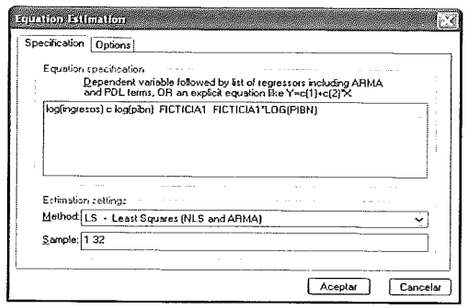


Figura 4-221

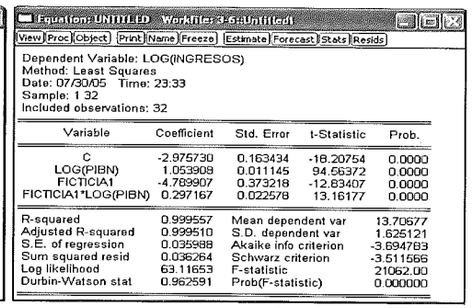


Figura 4-222

Se observa que ha mejorado la significatividad individual y conjunta de los parámetros y los coeficientes de determinación, y el valor del estadístico de Durbin Watson se acerca más a 2. En la tabla del estadístico de Durbin-Watson para $T = 32$, $k = 3$ y $\alpha = 0{,}05$ se tiene que $d_L = 1{,}244$ y $d_U = 1{,}65$. Como $d = 0{,}962591 < d_L = 1{,}244$, sigue existiendo autocorrelación positiva. Si representamos ahora los residuos contra la variable endógena del modelo utilizamos *Quick* → *Graph* → *Scatter*, rellenando la pantalla *Series List* como se indica en la Figura 4-223 y pulsando *OK*, se obtiene el gráfico de la Figura 4-224, cuya estructura ya es aleatoria.

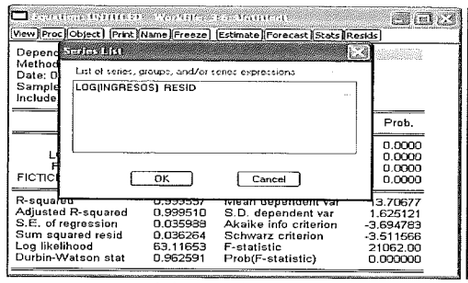


Figura 4-223

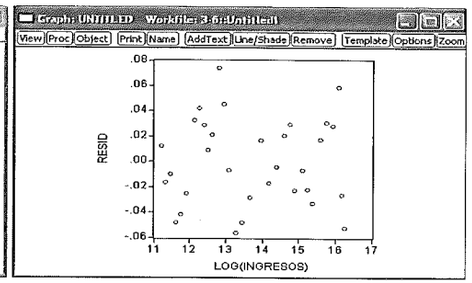


Figura 4-224

Ya sabemos que existe autocorrelación positiva. El siguiente paso será ***detectar el orden de esta autocorrelación***. Para ello puede utilizarse el correlograma residual. Desde la pantalla *Equation*, utilizamos *View → Residual Tests → Correlogram-Q-Statistics* (Figura 4-225). Se eligen 16 retardos (valor por defecto) para representar las funciones de autocorrelación (AC) y autocorrelación parcial (PAC) de los residuos y al pulsar *OK* se obtienen las gráficas de la Figura 4-226. La función de autocorrelación (correlograma) presenta una estructura de decrecimiento sinusoidal y la función de autocorrelación parcial sólo tiene el primer término significativamente distinto de cero (la única barra que atraviesa las bandas de confianza rayadas). Por lo tanto los residuos presentan una estructura autorregresiva de orden uno AR(1).

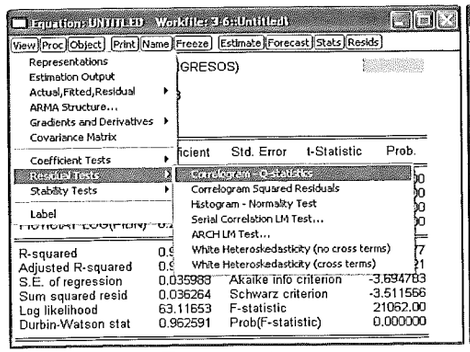


Figura 4-225

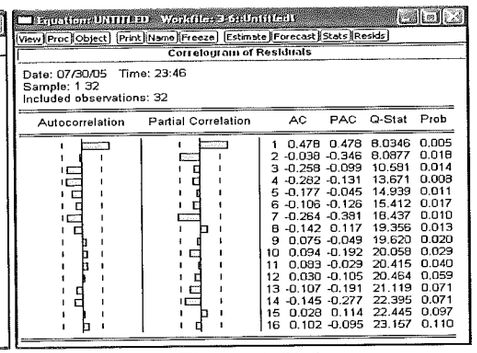


Figura 4-226

En presencia de autocorrelación se puede realizar la ***estimación mediante el método de Cochrane-Orcut con Eviews***. Se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification* (obsérvese la inclusión del término AR(1) como una variable explicativa más justificada por la presencia de autocorrelación de orden 1), se elige *Least Squares* en el campo *Method* para ajustar por mínimos cuadrados (Figura 4-227) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 4-228.

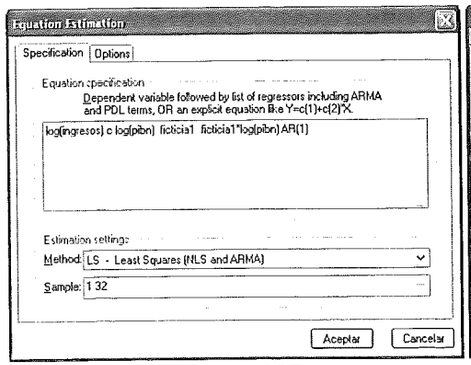


Figura 4-227

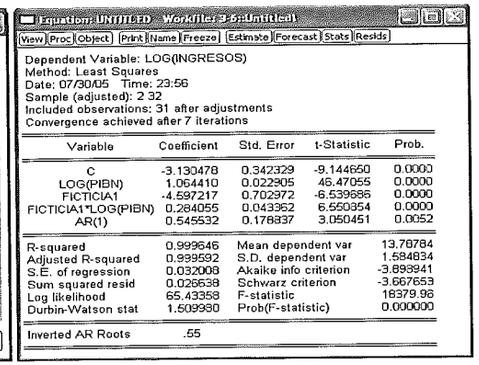


Figura 4-228

Se observa significatividad individual y conjunta de los parámetros muy alta y muy buen coeficiente de determinación. También es significativo el término autorregresivo AR(1). Por lo tanto, el modelo finalmente estimado tendrá la siguiente ecuación:

$$LOG(INGRESOS)_t = -3.130478116 + 1.064410042*LOG(PIBN)_t - 4.597217161*FICTICIA1_t + 0.2840553815*FICTICIA1*LOG(PIBN)_t + u_t$$

La estructura residual será la siguiente:

$$u_t = 0{,}545532*u_{t-1} + e_t$$

No hay problemas de ***normalidad de los residuos*** porque utilizando *View → Residual Tests → Histogram – Normality Test* desde la pantalla *Equation,* se obtienen los resultados de la Figura 4-229 que presenta un p-valor del estadístico de Jarque-Vera superior a 0,05. Tampoco hay problemas de ***heteroscedasticidad*** porque utilizando *View → Residual Tests → White Heteroskedasticity (no cross terms)* desde la pantalla *Equation,* se obtienen los resultados de la Figura 4-230 con p-valores de la *F* y la Chi-cuadrado superiores a 0,05.

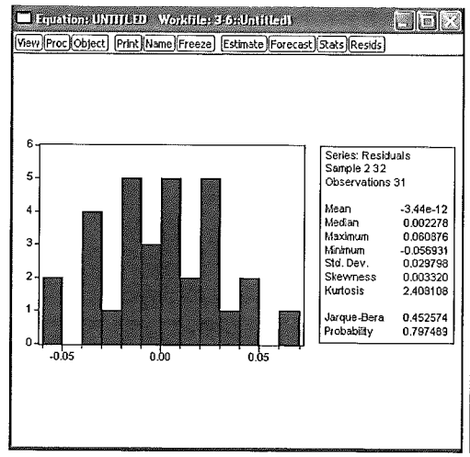


Figura 4-229

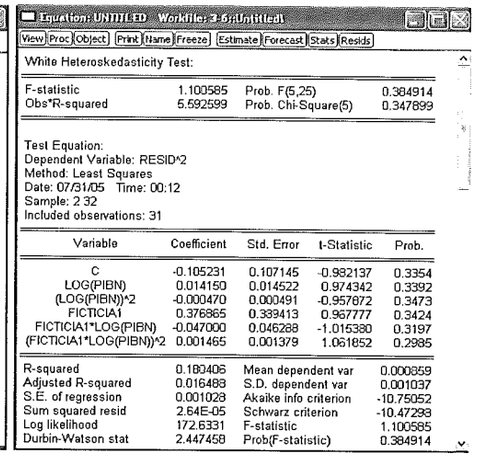


Figura 4-230

CAPÍTULO 5

MODELOS LOGIT, PROBIT, TOBIT, TRUNCADOS, RECUENTO, CENSURADOS Y DE SELECCIÓN MUESTRAL. HERRAMIENTAS

MODELOS DE VARIABLE DEPENDIENTE LIMITADA

La expresión funcional del modelo de análisis de la regresión múltiple es $y = F(x_1, x_2, \cdots, x_n)$. La regresión múltiple admite la posibilidad de trabajar con variables dependientes cuyo rango de valores está restringido (variables binarias con valores 0 y 1, variables con valores enteros positivos, etc.). En general, los modelos que admiten variables dependientes con rango restringido se denominan ***modelos de variable dependiente limitada***.

La mayoría de las variables económicas que se analizan presentan valores que están limitados de alguna manera, en muchas ocasiones porque deben ser positivos. Por ejemplo, el salario por hora, los precios de las viviendas, y los tipos de interés nominales deben ser mayores que cero. Pero no todas esas variables requieren un trato especial. No suele ser necesario ningún modelo econométrico especial para tratar las variables que son estrictamente positivas pero que toman muchos valores diferentes. Cuando la variable dependiente es discreta (***modelos de elección discreta***) y toma un reducido número de valores, no tiene sentido que la tratemos como si fuera una variable aproximadamente continua. El hecho de que la variable dependiente sea discreta no implica necesariamente que los modelos lineales no sean apropiados. Sin embargo, para respuestas binarias (***modelos de elección binaria***), suelen utilizarse los modelos Logit y Probit y en ciertos casos el ***modelo lineal de probabilidad***. También para respuestas múltiples (***modelos de elección múltiple***) se utilizan los modelo logit y probit.

Otro ejemplo importante de variable dependiente limitada es la variable de recuento, que toma valores enteros no negativos. Un ***modelo de datos de recuento*** es aquel que tiene como variable dependiente una variable discreta de recuento que toma valores enteros no negativos. Los *modelos de regresión de Poisson* son apropiados para analizar las variables de recuento.

En el análisis econométrico, aparecen otros tipos de variables dependientes limitadas, especialmente cuando analizamos el comportamiento de individuos, familias, o empresas. Por ejemplo, supongamos que durante un año concreto una gran cantidad de familias de una región no realiza ninguna donación caritativa. Por tanto, la distribución de la población de las donaciones que realizan las familias está muy dispersa sobre valores positivos, pero con una alta concentración en el valor cero. Estamos aquí ante un ejemplo de solución de esquina. El ***modelo Tobit*** está diseñado explícitamente para variables dependientes que presentan soluciones de esquina. Estas variables valen cero para una proporción de la población considerable pero se distribuye de forma aproximadamente continua para los valores positivos.

Otro tipo de variables dependientes limitadas son las originadas por la censura de datos. Un ***modelo de regresión censurado*** es aquel cuya variable dependiente está censurada por encima (censura superior) o por debajo (censura inferior) de algún valor, es decir, la variable dependiente no se observa para una parte de la población. En los modelos de regresión censurados la variable dependiente subyacente es aproximadadmente continua, pero está censurada inferior o superiormente, debido a la forma en que recopilamos los datos o a limitaciones institucionales. Un ***modelo de regresión truncada*** es parecido a un modelo de regresión censurada, pero tiene alguna diferencia importante: en un modelo de regresión truncada, no disponemos de los datos acerca de algún segmento importante de la población. Normalmente, esto ocurre cuando el objetivo de una muestra es un subconjunto concreto de la población y, quizá porque es costoso, ignora al resto de la población. Se trata de un caso particular del *problema general de la selección muestral*, donde observamos una muestra no aleatoria de la población subyacente.

MODELOS DE ELECCIÓN DISCRETA

La expresión funcional del modelo de análisis de la regresión múltiple es $y = F(x_1, x_2, \cdots, x_n)$. La regresión múltiple admite la posibilidad de trabajar con variables dependientes discretas en vez de continuas para permitir la modelización de fenómenos discretos. Cuando la variable dependiente es una variable discreta que refleja decisiones individuales en las que el conjunto de elección está formado por alternativas separadas y mutuamente excluyentes estamos ante los ***modelos de elección discreta***. Cuando la variable dependiente es discreta y toma sólo un número pequeño de valores no tiene sentido tratarla como si fuera una variable continua y suele interesar *caracterizar la probabilidad de que un agente tome una determinada decisión discreta*, condicional a los valores de ciertas variables explicativas. Estas funciones de distribución que caracterizan probabilidades para cada valor de las variables explicativas suelen ser no lineales y no suelen tener solución analítica por lo que suele ser necesario recurrir a métodos numéricos. Los modelos de elección discreta en los que el conjunto de elección tiene sólo dos alternativas posibles se llaman ***modelos de elección binaria***. Cuando el conjunto de elección tiene varios valores discretos nos encontramos ante los ***modelos de elección múltiple o modelos multinomiales***.

Los modelos de elección discreta se denominan *modelos de datos de recuento* cuando los valores de la variable dependiente discreta son números que no reflejan categorías. En caso de que los valores numéricos de la variable dependiente discreta reflejen categorías, los modelos se denominan *modelo de elección discreta categóricos*, y suelen clasificarse en *modelos de elección discreta categóricos ordenados* (los valores numéricos no tienen significado cuantitativo y reflejan un orden de categorías) y *modelos de elección discreta categóricos no ordenados* (los valores numéricos reflejan únicamente categorías).

MODELOS DE ELECCIÓN DISCRETA BINARIA

Dentro de los ***modelos de elección discreta*** en los que el conjunto de elección tiene sólo dos alternativas posibles mutuamente excluyentes, consideraremos el modelo lineal de probabilidad, el modelo Logit y el modelo Probit.

Modelo lineal de probabilidad

Partimos del modelo de regresión lineal habitual:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

una de cuyas hipótesis es:

$$E(\varepsilon | X_1, X_2, \ldots, X_k) = 0$$

lo que nos lleva a escribir el modelo como:

$$E(Y | X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Pero en el caso de los modelos de elección discreta en los que el conjunto de elección tiene sólo dos alternativas posibles mutuamente excluyentes, Y es una variable aleatoria de Bernouilli de parámetro p, lo que nos permite escribir:

$$E(Y | X_1, \ldots, X_k) = P(Y = 1 | X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Estamos ahora ante el ***modelo lineal de probabilidad***, donde, por ejemplo, β_1 mide la variación en la probabilidad de "éxito" ($Y = 1$) ante una variación unitaria en X_1 (con todo lo demás constante).

Como Y es una variable aleatoria de Bernouilli:

$$V(Y|X_1,\ldots,X_k) = P(Y=1|X_1,\ldots,X_k)\big(1-P(Y=1|X_1,\ldots,X_k)\big)$$

Tenemos entonces:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u \Rightarrow u = Y - \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

$$V(u) = V(Y - \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) = V(Y \mid X_1,\ldots,X_k)$$

para cada observación $V(u_i) = p_i(1\text{-}p_i)$ ya que Y es una variable aleatoria de Bernouilli.

Estamos entonces ante un modelo con heteroscedasticidad porque la varianza del error no es constante, ya que para cada valor de $X_1,\ldots,X_k$, la varianza del error tiene un valor diferente ($V(u)$ no constante). Además, Y es una variable de Bernouilli, con lo que tampoco se cumple la hipótesis de normalidad. Ello obliga a estimar estos modelos por un método alternativo a mínimos cuadrados ordinarios, por ejemplo, utilizando estimadores máximo verosímiles o mínimos cuadrados generalizados.

Realizada la estimación del modelo lineal de probabilidad tenemos que:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k = \hat{P}$$

se puede interpretar como una estimación de la probabilidad de "éxito" (de que $Y = 1$). En algunas aplicaciones tiene sentido interpretar $\hat{\beta}_0$ como la probabilidad de éxito cuando todas las X_j valen 0.

Otra limitación importante del modelo lineal de probabilidad es que para ciertas combinaciones de las variables explicativas $X_1,\ldots,X_k$, las probabilidades estimadas pueden ser mayores que cero o menores que uno.

Modelos Probit y Logit

Podemos considerar los ***modelos Logit y Probit*** como modelos de respuesta binaria:

$$P(Y=1|X_1,X_2,\ldots,X_k) = G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$

que, para evitar los problemas del modelo lineal de probabilidad, se especifican como $Y = G(X\beta)$, donde G es una función que toma valores estrictamente entre 0 y 1 ($0<G(Z)<1$) para todos los números reales z. Según las diferentes definiciones de G tenemos los distintos modelos de elección binaria.

Si $G(z) = \dfrac{e^z}{1+e^z}$ estamos ante el ***modelo Logit***, cuya expresión será:

$$Y = G(z) = G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}$$

En el caso del ***modelo Probit*** tenemos:

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v)\mathbf{d}v$$

donde $\Phi(z) = \dfrac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$ es la función de densidad de la *normal* (0,1).

La expresión del modelo Probit será:

$$Y = G(z) = G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) = \int_{-\infty}^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k} \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}} dv$$

Los modelos Probit y Logit, como son modelos no lineales, no podremos estimar por MCO y tendremos que emplear métodos de máxima verosimilitud.

Supongamos que tenemos n observaciones idéntica e independientemente distribuidas (muestra aleatoria) que siguen el modelo:

$$P(Y=1|\mathbf{X}) = G(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)$$

Para obtener el estimador de máxima verosimilitud (MV), condicionado a las variables explicativas, necesitamos la función de verosimilitud:

$$L(\beta) = \prod_{Y_i=1} P_i \prod_{Y_i=0} (1-P_i) = \prod_{i=1}^{n} G(X_i'\beta)^{Y_i}\big(1-G(X_i'\beta)\big)^{1-Y_i}$$

con:

$$P_i = P(Y_i = 1|X_{1i},\ldots,X_{ki}) = G(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}) = G(X_i'\beta)$$

El estimador de MV de β es el que maximiza el logaritmo de la función de verosimilitud:

$$l(\beta) = \ln L(\beta) = \sum_{i=1}^{n} \left[Y_i \ln G(X_i'\beta) + (1-Y_i)\ln\big(1-G(X_i'\beta)\big)\right]$$

que será un estimador consistente, asintóticamente normal y asintóticamente eficiente.

Las condiciones de primer orden serán:

$$S(\beta)=\sum_{i=1}^{n}\left[\frac{Y_i}{G(X_i'\beta)}-\frac{(1-Y_i)}{(1-G(X_i'\beta))}\right]X_i g(X_i'\beta)=$$

$$=\sum_{i=1}^{n}\left[\frac{Y_i-G(X_i'\beta)}{G(X_i'\beta)(1-G(X_i'\beta))}\right]X_i g(X_i'\beta)=0$$

donde $g(.)$ es la función de densidad de la normal o la logística (derivada de la función de distribución).

La no linealidad del problema hace que para obtener el estimador MV de β necesitemos aplicar un algoritmo iterativo y obtener el estimador por métodos numéricos iterativos. Mediante el algoritmo Scoring tenemos:

$$\hat{\beta}^{k+1}=\hat{\beta}^{k}+\left[I(\hat{\beta}^{k})\right]^{-1}S(\hat{\beta}^{k})$$

La matriz de covarianzas asintótica de $\hat{\beta}$ se estima como:

$$A\hat{\text{var}}(\hat{\beta})=\left[I(\hat{\beta})\right]^{-1}=\left(\sum_{i=1}^{n}\frac{\left[g(X_i'\hat{\beta})\right]^2 X_i X_i'}{G(X_i'\hat{\beta})\left(1-G(X_i'\hat{\beta})\right)}\right)^{-1}$$

Para realizar ***contrastes de hipótesis en los modelos Logit y Probit*** tendremos en cuenta que la raíz cuadrada de los elementos de la diagonal principal de la matriz de covarianzas asintótica son los errores estándar (asintóticos) de cada uno de los $\hat{\beta}_j$, que los podemos emplear para construir los estadísticos t (que tendrán una distribución asintótica normal) o intervalos de confianza aproximados para cada parámetro. También podemos contrastar varias restricciones simultáneamente. Lo habitual es que lo que nos interese sean restricciones de exclusión por lo que es en lo que nos vamos a centrar.

Para contrastar la hipótesis nula de que un conjunto de parámetros es igual a cero podemos emplear varios procedimientos:

- ***Estadístico de Wald.*** Se distribuye asintóticamente como una *Chi-cuadrado* con q (nº de restricciones) grados de libertad y lo proporcionan la mayoría de los programas.

- ***Contraste de razón de verosimilitudes (Likelihood Ratio (LR) test).*** Se basa en la diferencia entre el logaritmo de la función de verosimilitud en el modelo sin restringir y en el restringido:

$$LR=2\left(l(\hat{\beta}_{NR})-l(\hat{\beta}_{R})\right)$$

que se distribuye asintóticamente como una *Chi-cuadrado* con q grados de libertad.

En cuanto a las ***medidas de la bondad de ajuste en los modelos Logit y Probit*** tenemos:

- ***Porcentaje de predicciones correctas.*** Para cada i calculamos la probabilidad estimada de que $Y_i = 1$:

$$\hat{P}_i=\hat{P}(Y_i=1|X_{1i},\ldots,X_{ki})=G(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\ldots+\hat{\beta}_kX_{ki})$$

- Si $\hat{P}_i>0{,}5$ nuestra predicción será que Y_i es 1 y si $\hat{P}_i\leq 0{,}5$ nuestra predicción será que Y_i es 0. El % de veces en que el valor de Y_i observado coincida con la predicción es el % de predicciones correctas. Lo interesante es calcular por separado el % de predicciones correctas de ceros y de unos.

- ***Pseudo – R^2 (de McFadden).*** Está basado en el logaritmo de la función de verosimilitud:

$$Pseudo-R^2=1-\frac{l(\hat{\beta})}{l(\hat{\beta}_0)}$$

donde $l(\hat{\beta})$ es el logaritmo de la función de verosimilitud para el modelo estimado y $l(\hat{\beta}_0)$ el de un modelo sólo con término constante. Como $|l(\hat{\beta})| < |l(\hat{\beta}_0)|$, el valor $Pseudo-R^2$ está entre 0 y 1.

- ***Criterios de Información.*** Son medidas que tratan de buscar un equilibrio entre la bondad del ajuste, medida en base al valor del logaritmo de la función de verosimilitud, y una especificación parsimoniosa del modelo. (Ejemplos: *Akaike* (AIC), *Schwarz* (SC) y *Hannan-Quinn* (HQ)). Se escoge el modelo con menor valor del criterio de información.

A la hora de ***interpretar las estimaciones en los modelos Probit y Logit***, generalmente lo que nos interesa es conocer el efecto de variaciones en una variable X_j sobre la probabilidad de respuesta, que si la variable es continua será:

$$\Delta\hat{P}(Y=1|\mathbf{X}) \approx \left[g(\mathbf{X}\hat{\beta})\hat{\beta}_j\right]\Delta X_j$$

Como $g(X\hat{\beta})$ depende de X habrá que calcular los efectos parciales para valores interesantes de X (las medias muestrales, valores máximos y mínimos de las variables de interés, etc.). También se puede calcular el efecto parcial para cada individuo y después calcular su media.

El ***efecto parcial de una variable continua X_j sobre la probabilidad de respuesta*** $P(Y=1|X)$ será:

$$\frac{\partial P(Y=1|\mathbf{X})}{\partial X_j} = g(\mathbf{X}\beta)\beta_j$$

donde $g(.)$ es la función de densidad de la logística (*logit*) o de la normal estándar (*probit*). Este efecto varía de individuo a individuo. Como en el caso del Probit y del Logit, $g(z)>0$ para todo z, ***el signo del efecto parcial de X_j es el mismo que el de β_j.***

El ***efecto relativo de dos variables continuas X_j y X_h no depende de X.*** Nótese que el cociente de los efectos parciales es β_j/β_h.

Si X_1, por ejemplo, es una variable explicativa ficticia, el efecto parcial de que varíe de 1 a 0 vendrá dado por:

$$G(\beta_0+\beta_1+\beta_2X_2+\ldots+\beta_kX_k) - G(\beta_0+\beta_2X_2+\ldots+\beta_kX_k)$$

que también varía de un individuo a otro, pues depende de los valores de todas las X_j.

Como en el Probit $g(0) \approx 0{,}4$, en el Logit $g(0) \approx 0{,}25$ y en el MPL $g(0) = 1$, se puede obtener la siguiente relación entre las estimaciones:

$$\hat{\beta}_{Logit} \approx 1{,}6\hat{\beta}_{Probit} \qquad \hat{\beta}_{Logit} \approx 4\hat{\beta}_{MPL}$$

MODELOS DE ELECCIÓN MÚLTIPLE

Los modelos estudiados hasta ahora son modelos de elección discreta en los que el conjunto de elección tiene sólo dos alternativas posibles y que se llaman *modelos de elección binaria*. Pero cuando el conjunto de elección tiene varios valores discretos nos encontramos ante los ***modelos de elección múltiple o modelos multinomiales***. Estudiaremos a continuación los más habituales.

Modelo Logit Multinomial

El ***Modelo Logit Multinomial*** es una extensión del modelo binario para el caso en el que la respuesta, "desordenada", tiene más de 2 posibilidades. Sea (X_i, Y_i) una muestra aleatoria de la población ($i = 1..n$).

Al igual que en el caso binario, lo que nos interesa es saber cómo afectan los cambios en los elementos de X a las probabilidades de respuesta:

$$P(Y=j|X_1,X_2,\ldots,X_k) = P(Y=j|\mathbf{X}) \qquad j=0,1,\ldots J$$

En el Modelo Logit Multinomial las probabilidades de respuesta son:

$$P(Y=j|\mathbf{X}) = \frac{\exp(\mathbf{X}\beta_j)}{1+\sum_{h=1}^{J}\exp(\mathbf{X}\beta_h)} = p_j(\mathbf{X},\beta) \qquad j=1,\ldots J$$

$$P(Y=0|\mathbf{X}) = \frac{1}{1+\sum_{h=1}^{J}\exp(\mathbf{X}\beta_h)} = p_0(\mathbf{X},\beta)$$

Si J=1, estamos en el caso binario.

En estos modelos los efectos parciales son complicados y ni siquiera el signo del parámetro nos da el signo del efecto. Si X_k es continua, el efecto parcial será:

$$\frac{\partial P(Y=j|\mathbf{X})}{\partial X_k} = P(Y=j|\mathbf{X})\left\{\beta_{jk} - \left[\sum_{h=1}^{J}\beta_{hk}\exp(\mathbf{X}\beta_h)\right]/\,g(\mathbf{X},\beta)\right\}$$

donde β_{hk} es el elemento k-ésimo de β_h y:

$$g(\mathbf{X},\beta) = 1+\sum_{h=1}^{J}\exp(\mathbf{X}\beta_h)$$

Se observa que:

$$\frac{P(Y=j|\mathbf{X})}{P(Y=0|\mathbf{X})} = \frac{p_j(\mathbf{X},\beta)}{p_0(\mathbf{X},\beta)} = \exp(\mathbf{X}\beta_j)$$

luego:

$$\Delta \frac{p_j(\mathbf{X},\beta)}{p_0(\mathbf{X},\beta)} \approx \beta_{jk} \exp(\mathbf{X}\beta_j)\Delta X_k$$

Además:

$$\log\left(\frac{p_j(\mathbf{X},\beta)}{p_0(\mathbf{X},\beta)}\right) = \mathbf{X}\beta_j$$

y por tanto:

$$\Delta \log\left(\frac{p_j(\mathbf{X},\beta)}{p_0(\mathbf{X},\beta)}\right) \approx \beta_{jk}\Delta X_k$$

En general:

$$\log\left(\frac{p_j(\mathbf{X},\beta)}{p_h(\mathbf{X},\beta)}\right) = \mathbf{X}(\beta_j - \beta_h)$$

La probabilidad de elegir j, si la elección es entre j y h, sigue un modelo Logit estándar con vector de parámetros β_j - β_h :

$$P(Y = j|Y = j \text{ o } Y = h, \mathbf{X}) = \Lambda\left[\mathbf{X}(\beta_j - \beta_h)\right] = \frac{\exp\left[\mathbf{X}(\beta_j - \beta_h)\right]}{1 + \exp\left[\mathbf{X}(\beta_j - \beta_h)\right]}$$

El Modelo Logit Multinomial se estima por máxima verosimilitud. El logaritmo de la función de verosimilitud condicional viene dado por:

$$l(\beta) = \sum_{i=1}^{n} \sum_{j=0}^{J} 1[Y_i = j] \log[p_j(X_i, \beta)]$$

y en general obtendremos estimadores consistentes y asintóticamente normales.

Modelo Logit Condicional

El Modelo Logit Condicional se utiliza para problemas en los que las elecciones del individuo (o empresa) se realizan, al menos en parte, en base a los atributos observables de cada alternativa. En el Logit Multinomial, las variables explicativas no cambian entre alternativas, es decir, para cada i, X_i contiene variables específicas del individuo, no de las alternativas y es apropiado para situaciones donde las características de las alternativas no son importantes (o no las observamos). En el Modelo Logit Condicional, para cada i, X_i contiene variables específicas del individuo y de las alternativas y es apropiado para situaciones donde se observan las características alternativas.

Supongamos que disponemos de una muestra aleatoria de tamaño n de la población subyacente ($i = 1,\ldots, n$). La utilidad de elegir la alternativa j es:

$$Y_{ij}^* = X_{ij}\beta + a_{ij} \qquad j = 0,1,2,\ldots,J$$

donde a_{ij} son inobservables que afectan a los gustos y X_{ij} varía entre alternativas (y puede que también entre individuos) pero no puede contener elementos que sólo varíen entre individuos (no contiene término constante).

Como ejemplo podemos considerar el tiempo que tarda en llegar al trabajo el individuo i que emplea el medio de transporte j o el coste del transporte j (que puede o no variar entre individuos).

Si las perturbaciones a_{ij} se distribuyen independientemente con distribución Weibull con:

$$F(a) = \exp[-\exp(-a)]$$

entonces el Modelo Logit Condicional se expresa como sigue:

$$P(Y_i = j|X_i) = p_j(\mathbf{X}) = \frac{\exp(X_{ij}\beta)}{\sum_{h=0}^{J} \exp(X_{ih}\beta)} \qquad j = 0,1,\ldots,J$$

Los efectos marginales se expresan como sigue:

$$\frac{\partial p_j(\mathbf{X})}{\partial X_{jk}} = p_j(\mathbf{X})[1 - p_j(\mathbf{X})]\beta_k \qquad j = 0,1,\ldots,J, k = 1,\ldots,K$$

$$\frac{\partial p_j(\mathbf{X})}{\partial X_{hk}} = -p_j(\mathbf{X})p_h(\mathbf{X})\beta_k \qquad j \neq h, k = 1,\ldots,K$$

donde β_k es el elemento k-ésimo de β.

La estimación del Modelo Logit Condicional se realiza por el método de máxima verosimilitud.

En el Modelo Logit Condicional también se cumple el supuesto de *independencia de alternativas irrelevantes IAI* porque el cociente de probabilidades de elección entre 2 alternativas ($Y=j$ *en lugar de* $Y=h$) es independiente del resto de alternativas. Se tiene

$$\frac{p_j(\mathbf{X}_j)}{p_j(\mathbf{X}_h)} = \frac{\exp(\mathbf{X}_j\beta)}{\exp(\mathbf{X}_h\beta)} = \exp[(\mathbf{X}_j - \mathbf{X}_h)\beta]$$

El supuesto de *independencia de alternativas irrelevantes IAI*, que se deriva del supuesto de independencia de las perturbaciones, es conveniente para la estimación, pero puede suponer una limitación, ya que en algunas aplicaciones con alternativas similares es poco realista. El supuesto de *independencia de alternativas irrelevantes IAI* supone que la probabilidad relativa de dos alternativas existentes no se ve afectada por la inclusión de una tercera alternativa. Se puede contrastar el supuesto de IAI empleando el un contraste sugerido en Hausman y McFadden.

Modelo Logit Anidado

El Modelo Logit Anidado es un caso particular del Modelo Logit Condicional en el que se relaja el supuesto de IAI agrupando las J alternativas en L subgrupos de alternativas similares. Se emplea mucho para modelizar decisiones de consumo. Se sigue manteniendo la IAI dentro de cada grupo pero no entre grupos. Se considera el proceso de elección como que primero se elige uno de los L grupos $P(Y \in G_L|X)$ y después, en ese grupo, la alternativa j $P(Y = j|Y \in G_L, X)$. El objeto de interés $P(Y = j|\ X)$ se obtiene multiplicando las dos anteriores.

El Modelo Logit Anidado se puede estimar por el método de máxima verosimilitud con información completa (ambas decisiones) o en 2 etapas (Logit Condicional para la elección de cada j en cada grupo y después la probabilidad de elegir el grupo L).

Modelo Probit Multinomial

Cuando en un modelo Logit condicional se puede relajar el supuesto de IAI empleando modelos con supuestos más flexibles sobre a_{ij}, puede obtenerse el modelo Probit multinomial como un caso particular suyo.

Supongamos que a_{ij} sigue una distribución normal multivariante con correlaciones arbitrarias entre a_{ij} y a_{ih} (con $j \neq h$) estamos ante el ***Modelo Probit Multinomial***. Aunque este modelo es atractivo teóricamente, las probabilidades de respuesta son muy complicadas y la estimación máxima verosímil es casi imposible con más de 5 alternativas. Los avances econométricos recientes van haciendo más fácil el uso de estos modelos.

MODELOS LOGIT Y PROBIT ORDENADOS

Los modelos de elección múltiple vistos hasta ahora no tienen en cuenta la naturaleza ordinal de Y. A veces en los modelos de elección múltiple Y es una respuesta ordenada y el valor asignado a cada alternativa no es arbitrario. Estamos entonces ante los modelos de respuesta ordenada. Por ejemplo, cuando Y puede reflejar la valoración de un crédito es una escala de 0 a 6.

Sea Y una variable de respuesta ordenada que toma valores $\{0, 1, 2, \ldots, J\}$. El modelo Probit (o Logit) ordenado para Y (condicionado a unas variables explicativas X) se puede derivar de un modelo de variable latente:

$$Y^* = \mathbf{X}\beta + \varepsilon$$

donde X no contiene constante, β contiene k parámetros y $\varepsilon|X \rightarrow N(0,1)$. Sean $\alpha_1 < \alpha_2 < \ldots < \alpha_J$ puntos de corte (*threshold parameters*) desconocidos. Definimos:

$$\begin{aligned} &Y = 0 \quad si \quad Y^* \leq \alpha_1 \\ &Y = 1 \quad si \quad \alpha_1 < Y^* \leq \alpha_2 \\ &\ldots\ldots \\ &Y = J \quad si \quad Y^* > \alpha_J \end{aligned}$$

La distribución condicional de Y dado X vendrá dada por:

$$\begin{aligned} P(Y = 0|\mathbf{X}) &= P(Y^* \leq \alpha_1|\mathbf{X}) = P(\mathbf{X}\beta + \varepsilon \leq \alpha_1|\mathbf{X}) = \Phi(\alpha_1 - \mathbf{X}\beta) \\ P(Y = 1|\mathbf{X}) &= P(\alpha_1 < Y^* \leq \alpha_2|\mathbf{X}) = \Phi(\alpha_2 - \mathbf{X}\beta) - \Phi(\alpha_1 - \mathbf{X}\beta) \\ &\ldots\ldots \\ P(Y = J|\mathbf{X}) &= P(Y^* > \alpha_J|\mathbf{X}) = 1 - \Phi(\alpha_J - \mathbf{X}\beta) \end{aligned}$$

Si J=1 tenemos el Probit binario con la constante $-\alpha_1$ incluida dentro de $\Phi(.)$ (en los binarios solemos poner el punto de corte en cero y estimar la constante). Los parámetros α y β se pueden estimar por el método de máxima verosimilitud. Si en vez de emplear $\Phi(.)$ utilizamos la logística $\Lambda(.)$ tendremos el Modelo Logit Ordenado.

Para el Probit Ordenado tenemos que los efectos parciales son:

$$\frac{\partial p_0(\mathbf{X})}{\partial X_k} = -\beta_k \phi(\alpha_1 - \mathbf{X}\beta) \qquad \frac{\partial p_J(\mathbf{X})}{\partial X_k} = \beta_k \phi(\alpha_J - \mathbf{X}\beta)$$

$$\frac{\partial p_j(\mathbf{X})}{\partial X_k} = \beta_k[\phi(\alpha_{j-1} - \mathbf{X}\beta) - \phi(\alpha_j - \mathbf{X}\beta)] \quad 0 < j < J$$

El signo de β_k sólo determina el signo del efecto parcial para $P(Y=0|X)$ y $P(Y=J|X)$, pero no para el resto.

Podemos aplicar estos modelos de respuesta ordenada en casos en que Y tiene un sentido cuantitativo pero también nos interesa conocer la naturaleza de la respuesta ordenada discreta. En estos casos puede interesarnos conocer:

$$E(Y|\mathbf{X}) = a_0 P(Y = a_0|\mathbf{X}) + a_1 P(Y = a_1|\mathbf{X}) + \ldots + a_J P(Y = a_J|\mathbf{X})$$

donde a_0, a_1, ..., a_J son los valores que toma la variable. Una vez que estimemos las probabilidades podemos estimar $E(Y|X)$ para cualquier valor de X que nos interese.

MODELOS DE DATOS DE RECUENTO

Una tipología importante de variable dependiente limitada es la variable de recuento, que toma valores enteros no negativos. Un ***modelo de datos de recuento*** es aquel que tiene como variable dependiente una variable discreta de recuento que toma valores enteros no negativos. Los *modelos de regresión de Poisson* son apropiados para analizar las variables de recuento. También lo son los *modelos de regresión Exponencial* y los *modelos de regresión Binomial Negativa.*

Los modelos de datos de recuento se caracterizan porque no tienen, en general, un límite superior natural, toman valor cero para algunos miembros de la población y suelen tomar pocos valores.

Si Y es la variable de recuento y X_1, ..., X_k son las variables explicativas, normalmente estaremos interesados en:

$$E(Y|X_1,\ldots,X_k) = E(Y|\mathbf{X}).$$

En los casos en los que Y es estrictamente positiva podemos emplear la transformación logarítmica $log(Y)$ y usar el modelo lineal. Sin embargo, en los datos de recuento Y suele tomar valor cero para un porcentaje no despreciable de la población. Con datos de recuento lo que se suele hacer es modelizar $E(Y|\mathbf{X})$ eligiendo formas funcionales que aseguren valores positivos para todo X y todo valor de los parámetros.

Modelo de Regresión de Poisson

Para datos de recuento, en que la variable Y toma pocos valores, lo más habitual es asumir que Y dado X_1, ..., X_k sigue una distribución Poisson. La distribución Poisson viene completamente determinada por su media, con lo que nos vale con especificar $E(Y|\mathbf{X})$:

$$P(Y = h|\mathbf{X}) = \frac{\exp[-E(Y|\mathbf{X})][E(Y|\mathbf{X})]^h}{h!} \qquad h = 0,1,2\ldots$$

Una posibilidad que nos asegura valores positivos para todo valor de X y de los parámetros es modelizar la función esperanza condicional $E(Y|\mathbf{X})$ como una función exponencial:

$$E(Y|X_1,\ldots,X_k) = \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k) = \exp(\mathbf{X}\beta)$$

En este caso:

$$P(Y = h|\mathbf{X}) = \frac{\exp[-\exp(\mathbf{X}\beta)][\exp(\mathbf{X}\beta)]^h}{h!} \qquad h = 0,1,2\ldots$$

que nos permite calcular las probabilidades condicionadas.

Tomando logaritmos tenemos que:

$$\log[E(Y|\mathbf{X})] = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k = \mathbf{X}\beta$$

luego podemos decir que $100 \times \beta_j$ es aproximadamente la variación porcentual en $E(Y|\mathbf{X})$ cuando X_j varía en 1 unidad:

$$\%\Delta E(Y|\mathbf{X}) \approx 100\beta_j \Delta X_j$$

Podemos interpretar los coeficientes como si fueran un modelo lineal con variable dependiente en logaritmo.

Podemos medir la variación % exacta en $E(Y|\mathbf{X})$ ante una variación unitaria de X_k por $\exp(\beta_k) - 1$:

$$\%\Delta E(Y|\mathbf{X}) = \left(\frac{E(Y|\mathbf{X}+1)}{E(Y|\mathbf{X})} - 1\right) \times 100 = \left(\frac{\exp(\beta_0 + \beta_1 X + \beta_1)}{\exp(\beta_0 + \beta_1 X)} - 1\right) \times 100 =$$
$$= (\exp(\beta_1) - 1) \times 100$$

En base a los supuestos que hemos hecho sobre la distribución Poisson y sobre la forma de esperanza condicional, podemos construir el logaritmo de la función de verosimilitud como:

$$l(\beta) = \sum_{i=1}^{n} \{Y_i \mathbf{X}_i \beta - \exp(\mathbf{X}_i \beta)\}$$

donde se ha eliminado el término –log(Y_i!) porque no depende de β.

Maximizando esta función se obtiene el estimador MV de β, que si la distribución condicional de Y es Poisson y la $E(Y|\mathbf{X})$ está bien especificada será consistente, eficiente y asintóticamente normal. A partir de estas estimaciones se pueden obtener los errores estándar de los $\hat{\beta}_j$.

A veces la distribución Poisson impone restricciones que no se cumplen en las aplicaciones empíricas. En concreto, en la Poisson todas las probabilidades y momentos de orden superior están determinados por la media, por lo que $E(Y|X) = V(Y|X)$. Esta igualdad no se cumple en muchas aplicaciones. Sin embargo, aunque no se cumpla la distribución Poisson, seguiremos obteniendo estimadores consistentes y asintóticamente normales de los β_j si la media condicional está bien especificada.

Cuando Y dado $X_1, \ldots, X_k$ no sigue una distribución Poisson al estimador que se obtiene de maximizar el logaritmo de la función de verosimilitud:

$$l(\beta) = \sum_{i=1}^{n} \{Y_i \mathbf{X}_i \beta - \exp(\mathbf{X}_i \beta)\}$$

se le llama *estimador de cuasi máxima verosimilitud (CMV)*. Cuando estimamos por CMV si no se cumple el supuesto de $E(Y|X) = V(Y|X)$ hay que ajustar los errores estándar para que sean válidos, para realizar inferencia aunque la distribución condicional de Y esté mal especificada. Una posibilidad para ajustar los errores estándar, es suponer que la varianza es proporcional a la media:

$$V(Y|\mathbf{X}) = \sigma^2 E(Y|\mathbf{X})$$

donde $\sigma^2 > 0$ es un parámetro desconocido. Si $\sigma^2 = 1$ tenemos el *supuesto sobre la varianza de la Poisson.* Si $\sigma^2 > 1$ tenemos *sobredispersión*, que es lo que sucede en muchas aplicaciones. Si $\sigma^2 < 1$ tenemos *infradispersión*, que es raro en las aplicaciones empíricas.

Bajo el supuesto de varianza proporcional a la media es fácil ajustar los errores estándar de la Poisson obtenidos por máxima verosimilitud. Habrá que multiplicarlos por $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$, siendo $\hat{\sigma}^2$ un estimador consistente de σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \frac{\hat{u}_i^2}{\hat{Y}_i} \qquad \hat{u}_i = Y_i - \hat{Y}_i \qquad \hat{Y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k)$$

Los errores estándar así obtenidos se llaman *errores estándar GLM (Generalized Linear Models)*. Estos errores estándar están obtenidos bajo el supuesto de varianza proporcional a la media, pero también es posible obtener errores estándar para los estimadores de CMV del modelo Poisson sin restringir la varianza.

Bajo el supuesto de distribución Poisson, para realizar *contrastes de restricciones de exclusión,* podemos emplear el contraste de razón de verosimilitudes:

$$LR = 2\left(l(\hat{\beta}_{NR}) - l(\hat{\beta}_R)\right)$$

que se distribuye asintóticamente como una Chi-cuadrado con q grados de libertad.

Bajo el supuesto de varianza proporcional a la media, para realizar *contrastes de restricciones de exclusión*, basta con ajustar el contraste de Razón de verosimilitudes dividiéndolo por $\hat{\sigma}^2$ del modelo sin restringir (estadístico de CMV). Para medir la bondad del ajuste en estos modelos se puede emplear un R^2 definido como el cuadrado del coeficiente de correlación entre Y_i e $\hat{Y}_i$. Tiene la ventaja de que siempre estará entre 0 y 1.

Modelo de Regresión de Binomial Negativa

Existen otros modelos de regresión para datos de recuento empleando distribuciones que generalizan la Poisson, por ejemplo utilizando la distribución Binomial Negativa. Estamos entonces ante el modelo de regresión de Binomial Negativa. Este modelo se emplea para casos de sobredispersión ya que se supone que $V(Y|X) = \sigma^2 E(Y|X) = (1+\eta^2)\, E(Y|X)$. En este caso se estiman los parámetros β y η^2 conjuntamente por el método de máxima verosimilitud. Para que las estimaciones sean consistentes y eficientes es necesario que se cumpla el supuesto de binomial negativa. Si estimamos β para η^2 fijo, las estimaciones serán consistentes si la $E(Y|X)$ está bien especificada.

Modelo de Regresión Exponencial

En el Modelo de regresión exponencial se estiman los parámetro por máxima verosimilitud empleando la distribución exponencial. Si la $E(Y|X)$ está bien especificada los estimadores serán consistentes aunque la distribución no sea exponencial (como otros estimadores de CMV). Para obtener errores GLM se supone: $V(Y|X) = \sigma^2 [E(Y|X)]^2$.

Modelo de Regresión Normal

En el Modelo de regresión normal se estiman los parámetros por máxima verosimilitud empleando la distribución normal. Para σ^2 fijo, si la $E(Y|X)$ está bien especificada los estimadores serán consistentes aunque la distribución no sea normal.

MODELOS CENSURADOS: EL MODELO TOBIT

Un determinado tipo muy especial de variables dependientes limitadas son las originadas por la censura de datos. Un ***modelo de regresión censurado*** es aquel cuya variable dependiente está censurada por encima (censura superior) o por debajo (censura inferior) de algún valor, es decir, la variable dependiente no se observa para una parte de la población. En los modelos de regresión censurados la variable dependiente subyacente es aproximadamente continua, pero está censurada inferior o superiormente, debido a la forma en que recopilamos los datos o a limitaciones institucionales.

Un caso particular de censura de datos se presenta cuando la variable dependiente vale cero para una gran parte de la población y es continua para los valores positivos. Estamos entonces ante modelo de variable limitada con solución de esquina. El ***modelo Tobit*** está diseñado explícitamente para variables dependientes que presentan soluciones de esquina. Estas variables valen cero para una proporción de la población considerable pero se distribuye de forma aproximadamente continua para los valores positivos.

El modelo de regresión censurado Tobit o Tobit tipo I se expresa como sigue:

$$Y^* = \mathbf{X}\beta + u \qquad u|\mathbf{X} \sim N(0,\sigma^2) \quad Y = \max(0, Y^*)$$

que también puede escribirse como:

$$Y_i = \begin{cases} 0 & si \quad Y_i^* \le 0 \\ Y_i^* & si \quad Y_i^* > 0 \end{cases}$$

Hemos formulado el modelo Tobit en términos de una variable latente Y^*, que satisface los supuestos del modelo de regresión clásico, distribución normal, homocedástica y con media condicional lineal.

La densidad de Y dado X es la misma que la de Y^* dado X para los valores positivos. Además sabemos que:

$$P(Y=0|\mathbf{X}) = P(Y^* < 0|\mathbf{X}) = P(u < -\mathbf{X}\beta|\mathbf{X}) = P\left(\frac{u}{\sigma} < \frac{-\mathbf{X}\beta}{\sigma}|\mathbf{X}\right) =$$
$$= \Phi\left(\frac{-\mathbf{X}\beta}{\sigma}\right) = 1 - \Phi(\mathbf{X}\beta/\sigma)$$

Por tanto, dada una muestra aleatoria de la población, podemos escribir el logaritmo de la función de verosimilitud como:

$$l(\beta,\sigma) = \sum_{i=1}^{n} 1[Y_i = 0]\log[1 - \Phi(X_i\beta/\sigma)]$$
$$+ 1[Y_i > 0]\log\{(1/\sigma)\phi[(Y_i - X_i\beta)/\sigma]\}$$

Al maximizar la función de verosimilitud obtenemos los ***estimadores MV de β y de σ***. Se pueden obtener los errores estándar de los coeficientes estimados y construir estadísticos t de Student para contrastes. También se pueden emplear el estadístico de Wald y RV.

Interpretación de los coeficientes en el modelo Tobit

Observamos que:

$$E(Y|Y > 0, \mathbf{X}) = \mathbf{X}\beta + E(u|u > -\mathbf{X}\beta)$$

y utilizando que si $z \to N(0,1)$ $E(z|z>c) = \phi(c) / (1-\Phi(c))$ se tiene:

$$E(Y|Y > 0, \mathbf{X}) = \mathbf{X}\beta + \sigma E[(u/\sigma)|(u/\sigma) > -\mathbf{X}\beta/\sigma) =$$
$$= \mathbf{X}\beta + \sigma\phi(\mathbf{X}\beta/\sigma)/\Phi(\mathbf{X}\beta/\sigma) =$$
$$= \mathbf{X}\beta + \sigma\lambda(\mathbf{X}\beta/\sigma)$$

$$\lambda(X\beta/\sigma) = \frac{\phi(X\beta/\sigma)}{\Phi(X\beta/\sigma)} = \text{Ratio de Mills}$$

Entonces, si realizamos estimaciones del modelo Tobit por MCO empleando sólo las observaciones para las que $Y>0$, omitiendo los ceros, obtendríamos un estimador de β insesgado e inconsistente, porque omitiríamos el ratio de Mills que está correlacionado con X.

En el modelo Tobit tenemos:

$$P(Y=0|\mathbf{X}) = 1-\Phi(\mathbf{X}\beta/\sigma)$$

Por otro lado:

$$E(Y|\mathbf{X}) = P(Y=0|\mathbf{X})0 + P(Y>0|\mathbf{X})E(Y|Y>0,\mathbf{X}) =$$
$$= \Phi(\mathbf{X}\beta/\sigma)E(Y|Y>0,\mathbf{X})$$

Y teniendo presente el valor del segundo término de la ecuación anterior en función del ratio de Mills, podemos escribir:

$$E(Y|\mathbf{X}) = \Phi(\mathbf{X}\beta/\sigma)[\mathbf{X}\beta+\sigma\lambda(\mathbf{X}\beta/\sigma)] =$$
$$= \Phi(\mathbf{X}\beta/\sigma)\mathbf{X}\beta+\sigma\phi(\mathbf{X}\beta/\sigma)$$

Estamos entonces ante una función no lineal de X y de β, que nos permite llegar a la conclusión de que si estimamos el modelo Tobit por MCO como un modelo lineal con todas las observaciones no tendremos estimadores consistentes.

Efectos parciales en el modelo Tobit

Si X_j es una variable continua, el efecto parcial sobre el valor esperado de Y en la subpoblación de valores positivos dependerá de β_j y de un término entre 0 y 1 y valdrá (*):

$$\frac{\delta E(Y|Y>0,\mathbf{X})}{\delta X_j} = \beta_j\{1-\lambda(\mathbf{X}\beta/\sigma)[\mathbf{X}\beta/\sigma+\lambda(\mathbf{X}\beta/\sigma)]\}$$

Podemos estimar este efecto parcial empleando las estimaciones MV de β y de σ para los valores medios de X o para otros que sean de interés en nuestras aplicaciones. Si X_j es una variable binaria podemos calcular la diferencia entre $E(Y|Y>0,X)$ para $X_j = 1$ y $X_j = 0$. De forma similar se puede hacer para cualquier X_j discreta.

Así mismo, si X_j es una variable continua, podremos calcular el efecto parcial sobre el valor esperado de Y. Este efecto parcial se puede descomponer en dos partes, el efecto sobre el valor medio de Y en la parte positiva de la distribución, y el efecto sobre la probabilidad en la otra parte de la distribución. Tenemos (**):

$$\frac{\delta E(Y|\mathbf{X})}{\delta X_j} = \frac{\delta P(Y>0|\mathbf{X})}{\delta X_j}E(Y|Y>0,\mathbf{X}) + P(Y>0|\mathbf{X})\frac{\delta E(Y|Y>0,\mathbf{X})}{\delta X_j}$$

Como:

$$P(Y>0|\mathbf{X}) = \Phi(\mathbf{X}\beta/\sigma)$$

derivando se tiene (***):

$$\frac{\delta P(Y>0|\mathbf{X})}{\delta X_j} = (\beta_j/\sigma)\phi(\mathbf{X}\beta/\sigma)$$

Sustituyendo (*) y (***) en (**) se tiene:

$$\frac{\delta E(Y|\mathbf{X})}{\delta X_j} = \beta_j\Phi(\mathbf{X}\beta/\sigma)$$

Este valor de los efectos lo podemos estimar fácilmente empleando las estimaciones MV de β y de σ para los valores medios de X o para otros que sean de interés en nuestras aplicaciones.

El modelo Tobit está basado en los supuestos de normalidad y homocedasticidad en el modelo de la latente. Cuando no se cumplen, el problema no es sólo la inconsistencia del estimador MV de β y de σ. Ni siquiera las expresiones de las esperanzas condicionales son las adecuadas. Se puede generalizar el modelo para permitir ciertas formas de heterocedasticidad y de no normalidad. También existen contrastes de ciertos tipos de normalidad y de heterocedasticidad, como por ejemplo:

$$V(u|X) = \sigma^2\exp(Z\delta)$$

Pero, para incumplimientos leves de los supuestos, se piensa que el modelo Tobit puede generar una buena aproximación de los efectos parciales.

Modelo Tobit con censura en los datos

Hemos analizado el modelo Tobit para solución de esquina, pero podemos generalizar al caso en que haya censura de cualquier tipo en los datos. El caso más general es cuando la variable de interés es continua pero está censurada por encima (censura a la derecha) o por debajo de un valor (censura la izquierda) que no tiene porqué ser el cero.

Podemos escribir el modelo Tobit como:

$$Y' = \mathbf{X}\beta + u \qquad u|\mathbf{X}, c \sim N(0,\sigma^2) \quad Y = \min(Y', c)$$

que también puede escribirse como:

$$Y_i = \begin{cases} c & si \quad Y_i^* \geq c \\ Y_i^* & si \quad Y_i^* < c \end{cases}$$

Observamos Y^* o c (sabemos que Y^* es menor que un valor c pero no sabemos cuánto vale). Estamos considerando el modelo Tobit con censura a la derecha, pero si la censura fuese a la izquierda tendríamos:

$$Y_i = \begin{cases} c & si \quad Y_i^* \leq c \\ Y_i^* & si \quad Y_i^* > c \end{cases}$$ y con las dos censuras a la vez: $$Y_i = \begin{cases} c & si \quad Y_i^* \leq c \\ Y_i^* & si \quad c < Y_i^* \leq d \\ d & si \quad d < Y_i^* \end{cases}$$

Estimación máximo verosímil (MV) del modelo Tobit censurado

Si tuviéramos una muestra aleatoria podríamos estimar β por MCO y realizar la inferencia de forma habitual, pero la censura en los datos nos causa problemas de inconsistencia en la estimación MCO (con todas las observaciones o sin las censuradas). Dada una muestra aleatoria, podemos obtener estimaciones consistentes de β y de σ si estimamos por MV. Para las observaciones sin censura, $Y = Y^*$, la densidad de Y es la misma que la de Y^*. Para las observaciones censuradas tenemos:

$$P(Y = c|\mathbf{X}) = P(Y^* \geq c|\mathbf{X}) = P(u \geq c - \mathbf{X}\beta|\mathbf{X}) = 1 - \Phi[(c - \mathbf{X}\beta)/\sigma]$$

Calcularemos el logaritmo de la función de verosimilitud y realizaremos la estimación MV. En este modelo los β_j se interpretan como en el modelo de regresión lineal.

La expresión del logaritmo de la función de verosimilitud para una muestra de tamaño n es la siguiente:

$$LnMV = \sum_{i=1}^{c}\left(1 - \Phi\left(\frac{X_i\beta}{\sigma}\right)\right) + \sum_{i=c+1}^{n}\Phi\left(\frac{Y_i - X_i\beta}{\sigma}\right)$$

Estimación por mínimos cuadrados ordinarios en dos etapas del modelo Tobit censurado

Aunque la estimación del modelo Tobit censurado por MCO no puede hacerse directamente, sí podemos utilizar el siguiente método en dos etapas:

1. Se construye la variable artificial dicotómica que vale 1 para valores positivos de Y ($Y_i > 0$ o $Y_i > c$) y que vale 0 para valores de Y menores o iguales que cero ($Y_i \leq 0$ o $Y_i \leq c$). Se estima el modelo Probit correspondiente y a partir de los valores estimados se obtiene el ratio de Mills:

$$\lambda(X\beta/\sigma) = \frac{\phi(X\beta/\sigma)}{\Phi(X\beta/\sigma)}$$

2. Se estima por MCO el modelo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_k X_{ki} + \sigma\lambda_i + u_i$

SELECCIÓN MUESTRAL: MODELOS TRUNCADOS

Un ***modelo de regresión truncada*** es parecido a un modelo de regresión censurada, pero tiene alguna diferencia importante: en un modelo de regresión truncada, no disponemos de los datos acerca de algún segmento importante de la población. Normalmente, esto ocurre cuando el objetivo de una muestra es un subconjunto concreto de la población y, quizá porque es costoso, ignora al resto de la población. Se trata de un caso particular del ***problema general de la selección muestral***, donde observamos una muestra no aleatoria de la población subyacente.

Supongamos que tenemos un modelo poblacional subyacente que satisface los supuestos del modelo de regresión clásico:

$$Y = \mathbf{X}\beta + u \qquad u|\mathbf{X} \sim N(0,\sigma^2)$$

Si tuviéramos una muestra aleatoria de la población, el procedimiento de estimación más eficiente sería MCO. Sin embargo, en el caso del modelo truncado el problema es que no tenemos una muestra aleatoria, porque nuestra muestra observada presenta una acotación del tipo $Y_i \leq c_i$ o $Y_i > c_i$ (***Modelo Tobit truncado***). Si queremos estimar β y σ necesitamos conocer la distribución de Y_i dado X y que $Y_i \leq c_i$ o $Y_i > c_i$. Es decir:

$$g(Y|\mathbf{X}_i, c_i) = \frac{f(Y|\mathbf{X}_i\beta, \sigma^2)}{F(c_i|\mathbf{X}_i\beta, \sigma^2)}$$

$$Y_i \leq c_i \text{ o } Y_i > c_i$$

donde $f(Y|X_i\beta, \sigma^2)$ es la densidad de una normal con media $X_i\beta$ y varianza σ^2 y $F(c_i|X_i\beta, \sigma^2)$ es la función de distribución de una normal con media $X_i\beta$ y varianza σ evaluada en c_i.

Si dividimos la densidad de Y dado X entre la probabilidad de que $Y_i \leq c_i$ o $Y_i > c_i$ (dado X), tomamos logaritmos y sumamos para todo n maximizando el resultado obtenemos los estimadores MV de β y σ^2, que serán consistentes y asintóticamente normales. Pero si los supuestos de homocedasticidad y normalidad no se cumplen, los estimadores MV serán inconsistentes.

Estimación del modelo Tobit truncado

Supongamos que tenemos un modelo $Y_i = X_i\beta + u_i$ y que la variable Y_i está truncada a partir de una cierta cota o barrera a (sólo toma valores a partir de la barrera). Tenemos:

$$E(Y_i \mid X_i; Y_i > a) = X_i\beta + \sigma \frac{\phi\left(\frac{a - X_i\beta}{\sigma}\right)}{1 - \Phi\left(\frac{a - X_i\beta}{\sigma}\right)} = Z_i + \sigma\gamma(\alpha_i)$$

$$\gamma(\alpha_i) = \frac{\phi\left(\frac{a - X_i\beta}{\sigma}\right)}{1 - \Phi\left(\frac{a - X_i\beta}{\sigma}\right)} \qquad \alpha_i = \frac{a - X_i\beta}{\sigma}$$

El modelo de variable truncada se puede especificar como sigue:

$$(Y_i / Y_i > a) = X_i\beta + \sigma\gamma(\alpha_i) + u_i$$

Efectos parciales en el modelo Tobit truncado

Se trata de cuantificar el efecto de un incremento unitario de un regresor X_k sobre la variable explicada y que vendrá dado por la derivada parcial de su valor esperado respecto a la variable X_k tenemos:

$$\frac{\partial E(Y_i \mid X_i; Y_i > a)}{\partial X_k} = \beta_k + \sigma\frac{\gamma(\alpha_i)}{\partial X_k} = \beta_k + \sigma\frac{\partial\left[\frac{\phi\left(\frac{a - X_i\beta}{\sigma}\right)}{1 - \Phi\left(\frac{a - X_i\beta}{\sigma}\right)}\right]}{\partial X_k} =$$

$$= \beta_k - \sigma\left(\frac{\beta_k}{\sigma}\right) = \frac{\phi\left(\frac{a - X_i\beta}{\sigma}\right)}{1 - \Phi\left(\frac{a - X_i\beta}{\sigma}\right)}\left[\frac{\phi\left(\frac{a - X_i\beta}{\sigma}\right)}{1 - \Phi\left(\frac{a - X_i\beta}{\sigma}\right)} - \frac{a - X_i\beta}{\sigma}\right] =$$

$$= \beta_k - \beta_k[\{\gamma(\alpha_i)\}^2 - \alpha_i\gamma(\alpha_i)] = \beta_k[1 - [\{\gamma(\alpha_i)\}^2 - \alpha_i\gamma(\alpha_i)]] = \beta_k[1 - \delta(\alpha_i)]$$

Luego el efecto marginal de variable X_k respecto de la variable explicada equivale al coeficiente estimado β_k multiplicado por el factor de corrección $[1 - \delta(\alpha_i)]$ siendo:

$$\delta(\alpha_i) = \{\gamma(\alpha_i)\}^2 - \alpha_i\gamma(\alpha_i) \qquad \gamma(\alpha_i) = \frac{\phi\left(\frac{a - X_i\beta}{\sigma}\right)}{1 - \Phi\left(\frac{a - X_i\beta}{\sigma}\right)} \qquad \alpha_i = \frac{a - X_i\beta}{\sigma}$$

Estimación máximo verosímil (MV) del modelo Tobit truncado

La función de verosimilitud del modelo de regresión truncada se calcula a partir de la función de densidad:

$$f(Y_i \mid X_i; Y_i > a) = \frac{1}{\sigma}X_i\beta + \sigma\frac{\phi(Y_i)}{1 - \Phi(\alpha_i)} = Z_i + \sigma\gamma(\alpha_i)$$

Aplicando neperianos se obtiene que el logaritmo de la función de verosimilitud para una muestra de tamaño n es:

$$LnL = -\frac{n}{2}Ln2\pi - \frac{n}{2}Ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - X_i\beta)^2 - \sum_{i=1}^{n}Ln\left[1 - \Phi\left(\frac{a - X_i\beta}{\sigma}\right)\right]$$

Derivando e igualando a cero obtenemos los estimadores MV de β y de σ.

Estimación por el método de Amemiya en dos etapas

Aunque la estimación del modelo Tobit truncado por MCO no puede hacerse directamente, sí podemos utilizar el siguiente método en dos etapas:

1. Se estima por MCO el modelo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$ y se obtiene $\hat{Y}_i$

2. Se especifica la ecuación auxiliar $Y_i(Y_i - a) = \sigma^2 + \beta_0 Y_i + \beta_1 Y_i X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k Y_i X_{ki} + u_i$ y se estima por el método de las variables instrumentales utilizando como intrumentos los regresores $\{1, \beta_0, \beta_1, \beta_2, \ldots, \beta_k\}$.

CORRECCIÓN DE LA SELECCIÓN MUESTRAL: ESTIMACIÓN BIETÁPICA DE HECKMAN O HECKIT

La regresión truncada es un caso especial de un problema general conocido como ***selección muestral no aleatoria*** que aparece cuando la muestra de que se dispone para la estimación no es aleatoria.

Heckman ideó un procedimiento en dos etapas para estimar modelos en presencia de truncamiento incidental que salva el problema de la presencia de sesgo de selección:

1ª Etapa. Usando las n observaciones se estima un modelo Probit de s sobre Z a partir del cual obtenemos $\hat{\gamma}$. A partir de esas estimaciones calculamos para cada i $\hat{\lambda}_i = \lambda(Z_i\hat{\gamma})$.

2ª Etapa. Empleando la muestra seleccionada (s=1) estimamos por MCO la regresión de Y sobre X y $\hat{\lambda}_i$.

Mediante este procedimiento se obtienen estimadores consistentes y asintóticamente normales.

Un contraste de la existencia de sesgo de selección muestral, consiste en contrastar en la 2ª etapa la hipótesis de $\rho = 0$ empleando el estadístico t de $\hat{\lambda}_i$. Cuando $\rho \neq 0$ los errores estándar de la 2ª etapa no son válidos y hay que corregirlos.

Hemos supuesto que X es un subconjunto de Z ya que si excluimos inadecuadamente algún elemento de X podemos tener estimadores inconsistentes. Es importante que Z contenga al menos un elemento que no pertenezca a X. No es imprescindible para la identificación de β ya que para eso es suficiente con la no linealidad de $\hat{\lambda}_i$.

Pero si tenemos elementos adicionales, disminuye la correlación de $\hat{\lambda}_i$ con X y obtendremos estimadores de los β más precisos. Además, si en la primera etapa no incluimos elementos adicionales, la significatividad de $\hat{\lambda}_i$ en la 2ª etapa puede ser por selección muestral o por una mala especificación.

Además de empleando el procedimiento de 2 etapas, este tipo de modelos se pueden estimar por MV, aunque es más complicado porque necesitamos la distribución conjunta de Y y s.

SPSS Y LA REGRESIÓN LOGÍSTICA BINARIA

SPSS permite trabajar con modelos de variable dependiente limitada. Concretamente trata de modo bastante completo los modelos Logit y Probit incluidos los multinomiales.

Como ejemplo se considera el archivo *logitb.sav* que contiene datos de una muestra de 53 pacientes con cáncer de próstata en los que se mide la edad, el nivel de ácido que mide la extensión del tumor, el grado de agresividad del tumor, la etapa en la que se encuentra, los resultados de una radiografía y cuándo se ha detectado al intervenir quirúrgicamente que el cáncer se ha extendido a los nodos linfáticos. A partir de estos datos se trata de ajustar un modelo que permita predecir cuándo el cáncer se extiende a los nodos linfáticos (o no) sin necesidad de intervención quirúrgica.

SPSS incorpora el procedimiento ***Regresión logística binaria*** que implementa el análisis de este tipo de regresión logística. Ya sabemos que la regresión logística resulta útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de variables predictoras. Es similar a un modelo de regresión lineal pero está adaptado para modelos en los que la variable dependiente es dicotómica.

Los coeficientes de regresión logística pueden utilizarse para estimar la razón de las ventajas de la probabilidad de éxito a la probabilidad de fracaso (*odds ratio*) de cada variable independiente del modelo.

En cuanto a los datos, la variable dependiente debe ser dicotómica. Las variables independientes pueden estar a nivel de intervalo o ser categóricas. Si son categóricas, deben ser variables *dummy* o estar codificadas como indicadores (existe una opción en el procedimiento para recodificar automáticamente las variables categóricas).

En cuanto a supuestos, la regresión logística no se basa en supuestos distribucionales en el mismo sentido en que lo hace el análisis discriminante. Sin embargo, la solución puede ser más estable si los predictores tienen una distribución normal multivariante. Adicionalmente, al igual que con otras formas de regresión, la multicolinealidad entre los predictores puede llevar a estimaciones sesgadas y a errores típicos inflados. El procedimiento es más eficaz cuando la pertenencia a grupos es una variable categórica auténtica. Si la pertenencia al grupo se basa en valores de una variable continua (por ejemplo "CI alto" en contraposición a "CI bajo"), deberá considerarse el utilizar la regresión lineal para aprovechar la información mucho más rica ofrecida por la propia variable continua.

Para realizar un análisis de regresión logística binaria, elija en los menús *Analizar* → *Regresión* → *Logística binaria* (Figura 5-1) y seleccione las variables y las especificaciones para el análisis (Figura 5-2). Previamente es necesario cargar en memoria el fichero de nombre *logitb.sav* mediante *Archivo* → *Abrir* → *Datos*. La variable cualitativa dependiente va a ser *nodos* (negativo o positivo) y las variables independientes son *edad, grado, radigrafía, ácido* y *etapa*. Se ajustará un modelo que prediga la presencia o no de nodos linfáticos cancerígenos según los valores de las variables independientes.

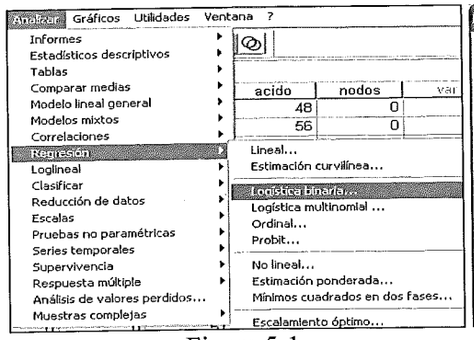


Figura 5-1

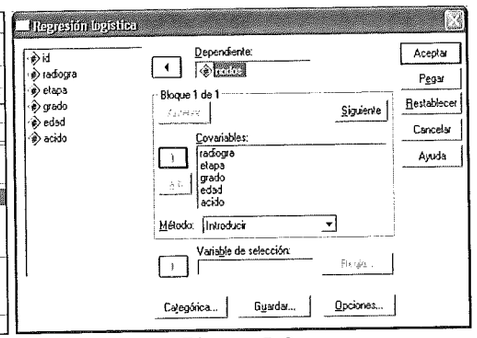


Figura 5-2

En los campos *Dependiente* y *Covariables* de la Figura 5-2 se introducen las variables dependiente e independientes del modelo. En el botón *Categórica* (Figura 5-3) puede especificar los detalles sobre cómo el procedimiento *Regresión logística* manipulará las variables categóricas. El campo *Covariables* contiene una lista de todas las covariables especificadas en el cuadro de diálogo principal. Si alguna de éstas son variables de cadena o son categóricas, sólo puede utilizarlas como covariables categóricas. En el campo *Covariables categóricas* se introduce la lista de las variables identificadas como categóricas. Cada variable incluye una notación entre paréntesis indicando el esquema de codificación de contraste que va a utilizarse. Las variables de cadena (señaladas con el símbolo < a continuación del nombre) estarán presentes ya en la lista *Covariables categóricas* por defecto. Seleccione cualquier otra covariable categórica de la lista *Covariables* y muévala a la lista *Covariables categóricas*. El campo *Variable de selección* permite limitar el análisis a un subconjunto de casos que tengan un valor particular en una variable.

El botón *Cambiar el contraste* le permite cambiar el método de contraste. Los métodos de contraste disponibles son: *Desviación* (cada categoría de la variable predictora, excepto la categoría de referencia, se compara con el efecto global, *Simple* (cada categoría de la variable predictora, excepto la misma categoría de referencia, se compara con la categoría de referencia), *Diferencia* o *contrastes de Helmert inversos* (cada categoría de la variable predictora, excepto la primera categoría, se compara con el efecto promedio de las categorías anteriores), *Helmert* (cada categoría de la variable predictora, excepto la última categoría, se compara con el efecto promedio de las categorías subsiguientes), *Repetida* (cada categoría de

la variable predictora, excepto la primera categoría, se compara con la categoría que la precede), *Polinómico* (contrastes polinómicos ortogonales en los que se supone que las categorías están espaciadas equidistantemente y sólo están disponibles para variables numéricas) e *Indicador* (los contrastes indican la presencia o ausencia de la pertenencia a una categoría y la categoría de referencia se representa en la matriz de contraste como una fila de ceros). Si selecciona *Desviación, Simple* o *Indicador*, elija *Primera* o *Última* como categoría de referencia. El método no cambia realmente hasta que se pulsa en *Cambiar*. Las covariables de cadena deben ser covariables categóricas. Para eliminar una variable de cadena de la lista *Covariables categóricas*, debe eliminar de la lista *Covariables* del cuadro de diálogo principal todos los términos que contengan la variable.

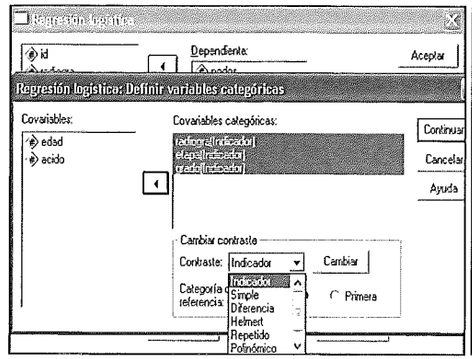


Figura 5-3

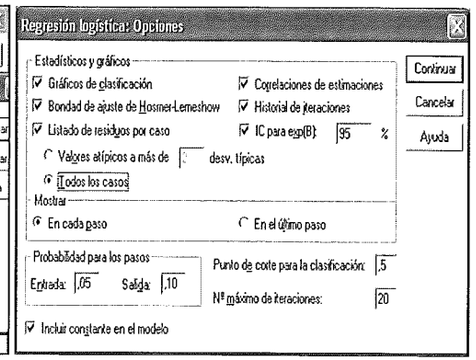


Figura 5-4

En el botón *Opciones* (Figura 5-4) puede especificar varias opciones para el análisis de regresión logística. La opción *Estadísticos y gráficos* le permite solicitar estadísticos y gráficos. Las opciones disponibles son *Gráficos de clasificación, Bondad de ajuste de Hosmer-Lemeshow, Listado de residuos por caso, Correlaciones de estimaciones, Historial de iteraciones* e *IC para exp(B)*. Seleccione una de las alternativas del grupo *Mostrar* para mostrar los estadísticos y los gráficos *En cada paso* o bien sólo para el modelo final, *En el último paso*. La opción *Probabilidad para el método por pasos* le permite controlar los criterios por los cuales las variables se introducen y se eliminan de la ecuación. Puede especificar criterios para la entrada o para la salida de variables. La opción *Punto de corte para la clasificación* le permite determinar el punto de corte para la clasificación de los casos. Los casos con valores pronosticados que han sobrepasado el punto de corte para la clasificación se clasifican como positivos, mientras que aquéllos con valores pronosticados menores que el punto de corte se clasifican como negativos.

Para cambiar los valores por defecto, introduzca un valor comprendido entre 0,01 y 0,99. La opción *Nº máximo de iteraciones* le permite cambiar el número máximo de veces que el modelo itera antes de finalizar. La opción *Incluir constante en el modelo* le permite indicar si el modelo debe incluir un término constante. Si se desactiva, el término constante será igual a 0.

La opción *Guardar* (Figura 5-5) permite guardar los resultados de la regresión logística como nuevas variables en el archivo de datos de trabajo. El campo *Valores pronosticados* guarda los valores pronosticados por el modelo. Las opciones disponibles son *Probabilidades* y *Grupo de pertenencia*. El campo *Influencia* guarda los valores de estadísticos que miden la influencia de los casos sobre los valores pronosticados. Las opciones disponibles son: *De Cook, Valores de influencia* y *DfBeta(s)*. El campo *Residuos* guarda los residuos. Las opciones disponibles son: *No tipificados, Logit, Método de Student, Tipificados* y *Desviación*. En todas las figuras el botón *Restablecer* permite restablecer todas las opciones por defecto del sistema y elimina del cuadro de diálogo todas las asignaciones hechas con las variables.

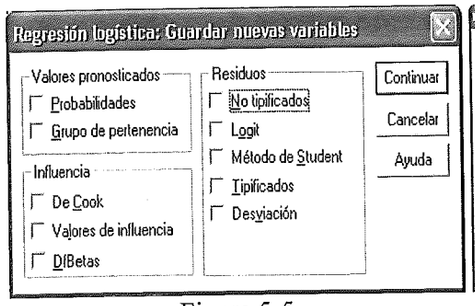


Figura 5-5

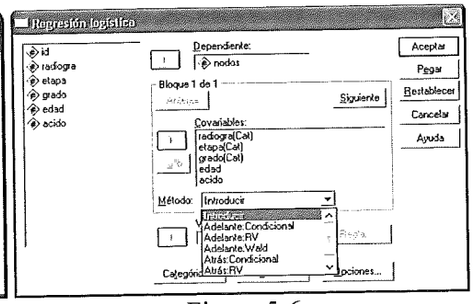
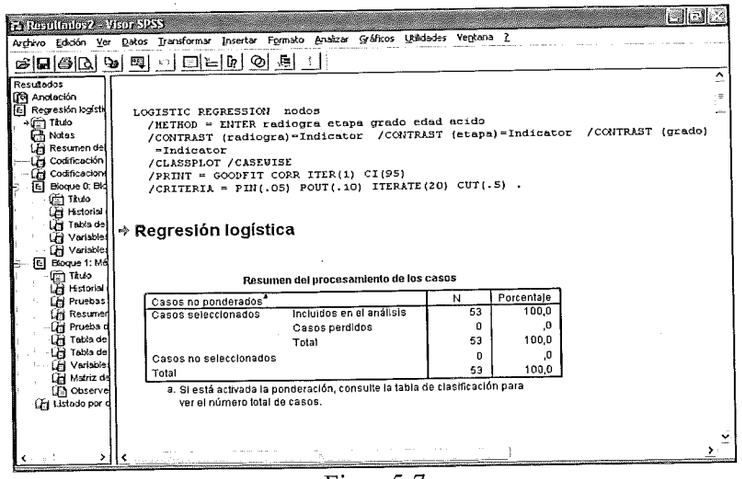


Figura 5-6

La Figura 5-6 presenta la pantalla del procedimiento *Regresión logística binaria* con todas las opciones ya incorporadas. Como *Método* de estimación puede usarse entrada en bloque de las variables (*Introducir*) o cualquiera de los métodos por pasos: Condicional, RV o Wald hacia delante y Condicional RV o Wald hacia atrás. Al hacer clic en *Aceptar*, se obtiene la salida del procedimiento (Figuras 5-7 a 5-16).



Figua 5-7

La Figura 5-8 presenta estadísticos de codificación de las variables. Las Figuras 5-9 y 5-10 presentan, para el Paso 0, el historial de iteraciones, la tabla de clasificación y las variables dentro y fuera de la ecuación con su significatividad y estadístico de Wald. Las Figuras 5-11 a 5-14 presentan, para el Paso 1, historial de iteraciones, estadísticos de la bondad de ajuste (pruebas omnibus con *Chi-cuadrado* del modelo, -2 log de la verosimilitud y R^2 de Cox y Snell y Nagelkerke), prueba de ajuste de Hosmer-Lemeshow con su tabla de contingencia, tabla de clasificación, variables en la ecuación y matriz de correlaciones. Los valores y p-valores para pruebas y estadísticos de ajuste indican buen ajuste y significatividad global e individual (variable a variable) del modelo. La Figura 5-15 presenta el gráfico de las probabilidades pronosticadas y grupos observados, y la Figura 5-16 presenta el listado por casos correspondiente incluidos residuos normales y tipificados.

Codificación de la variable dependiente

Valor original	Valor interno
Neg	0
Pos	1

Codificaciones de variables categóricas

		Frecuencia	Codificación de (1)
grado	0	33	1,000
	1	20	,000
etapa	0	26	1,000
	1	27	,000
radiogra	0	38	1,000
	1	15	,000

Figura 5-8

Bloque 0: Bloque inicial

Historial de iteraciones[a,b,c]

Iteración		-2 log de la verosimilitud	Coeficientes Constante
Paso 0	1	70,253	-,491
	2	70,252	-,501
	3	70,252	-,501

a. En el modelo se incluye una constante.
b. -2 log de la verosimilitud inicial: 70,252
c. La estimación ha finalizado en el número de iteración 3 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Tabla de clasificación[a,b]

Observado		Pronosticado nodos Neg	Pronosticado nodos Pos	Porcentaje correcto
Paso 0 nodos	Neg	33	0	100,0
	Pos	20	0	,0

Figura 5-9

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-,501	,283	3,123	1	,077	,606

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	radiogra(1)	11,283	1	,001
		etapa(1)	7,438	1	,006
		grado(1)	4,075	1	,044
		edad	1,094	1	,296
		acido	3,117	1	,077
	Estadísticos globales		19,451	5	,002

Figura 5-10

Bloque 1: Método = Introducir

Historial de iteraciones[a,b,c,d]

Iteración		-2 log de la verosimilitud	Coeficientes Constante	radiogra(1)	etapa(1)	grado(1)	edad	acido
Paso 1	1	49,605	2,969	-1,517	-1,025	-,472	-,045	,016
	2	48,193	4,146	-1,931	-1,448	-,694	-,065	,023
	3	48,126	4,419	-2,039	-1,558	-,758	-,069	,024
	4	48,126	4,433	-2,045	-1,564	-,761	-,069	,024
	5	48,126	4,433	-2,045	-1,564	-,761	-,069	,024

a. Método: Introducir
b. En el modelo se incluye una constante.
c. -2 log de la verosimilitud inicial: 70,252
d. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Figura 5-11

Pruebas ómnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	22,126	5	,000
	Bloque	22,126	5	,000
	Modelo	22,126	5	,000

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	48,126[a]	,341	,465

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	5,954	8	,652

Figura 5-12

Tabla de contingencias para la prueba de Hosmer y Lemeshow

		nodos = Neg		nodos = Pos		Total
		Observado	Esperado	Observado	Esperado	
Paso 1	1	5	4,807	0	,193	5
	2	5	4,659	0	,341	5
	3	5	4,441	0	,559	5
	4	3	4,185	2	,815	5
	5	3	3,907	2	1,093	5
	6	3	3,473	2	1,527	5
	7	4	2,913	1	2,087	5
	8	3	2,357	2	2,643	5
	9	1	1,429	4	3,571	5
	10	1	,830	7	7,170	8

Tabla de clasificación[a]

Observado			Pronosticado		
			nodos		Porcentaje correcto
			Neg	Pos	
Paso 1	nodos	Neg	28	5	84,8
		Pos	7	13	65,0
	Porcentaje global				77,4

a. El valor de corte es ,500

Figura 5-13

La tabla de clasificación de la Figura 5-13 indica que el modelo es bueno para clasificar en el futuro (predecir) ya que si se aplica la clasificación a las observaciones ya conocidas, se obtiene un porcentaje de éxitos del 77,4%. Esto hace suponer que para clasificaciones futuras se mantendrá el mismo porcentaje de éxitos. La lista por casos de la Figura 5-16 especifica qué observaciones se han clasificado bien y mal con el modelo.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B) Inferior	Superior
Paso 1[a]	radiogra(1)	-2,045	,807	6,421	1	,011	,129	,027	,629
	etapa(1)	-1,564	,774	4,084	1	,043	,209	,046	,954
	grado(1)	-,761	,771	,976	1	,323	,467	,103	2,115
	edad	-,069	,058	1,432	1	,231	,933	,833	1,045
	acido	,024	,013	3,423	1	,064	1,025	,999	1,051
	Constante	4,433	3,531	1,576	1	,209	84,155		

a. Variable(s) introducida(s) en el paso 1: radiogra, etapa, grado, edad, acido.

Matriz de correlaciones

		Constant	radiogra(1)	etapa(1)	grado(1)	edad	acido
Paso 1	Constant	1,000	-,239	-,092	-,070	-,946	-,082
	radiogra(1)	-,239	1,000	,126	-,003	,082	-,055
	etapa(1)	-,092	,126	1,000	-,212	,058	-,172
	grado(1)	-,070	-,003	-,212	1,000	,015	-,202
	edad	-,946	,082	,058	,015	1,000	-,142
	acido	-,082	-,055	-,172	-,202	-,142	1,000

Figura 5-14


```
                Step number: 1

                Observed Groups and Predicted Probabilities

         4 ⇕  NN                                                    ⇕
           ⇔  NN                                                    ⇔
           ⇔  NN                                                    ⇔
F          ⇔  NN                                                    ⇔
R        3 ⇕  NN    P          P                             P      ⇕
E          ⇔  NN    P          P                             P      ⇔
Q          ⇔  NN    P          P                             P      ⇔
U          ⇔  NN    P          P                             P      ⇔
E        2 ⇕  NNNN  N    PP     N        P    PP            P    P ⇕
N          ⇔  NNNN  N    PP     N        P    PP            P    P ⇔
C          ⇔  NNNN  N    PP     N        P    PP            P    P ⇔
Y          ⇔  NNNN  N    PP     N        P    PP            P    P ⇔
         1 ⇕  NNNNNNNPNNNN NNPNN  N    NN NNN N      P P  P P P NP PP ⇕
           ⇔  NNNNNNNPNNNN NNPNN  N    NN NNN N      P P  P P P NP PP ⇔
           ⇔  NNNNNNNPNNNN NNPNN  N    NN NNN N      P P  P P P NP PP ⇔
           ⇔  NNNNNNNPNNNN NNPNN  N    NN NNN N      P P  P P P NP PP ⇔
Predicted
⇩⇩⇩
```

26	S	P**	,139	N	,861	2,485
27	S	N	,255	N	-,255	-,585
28	S	N	,201	N	-,201	-,501
29	S	N	,304	N	-,304	-,661
30	S	N	,146	N	-,146	-,414
31	S	N**	,531	P	-,531	-1,064
32	S	N	,321	N	-,321	-,688
33	S	P	,695	P	,305	,663
34	S	P	,849	P	,151	,422
35	S	P**	,157	N	,843	2,320
36	S	N	,280	N	-,280	-,623
37	S	N**	,868	P	-,868	-2,560
38	S	N	,471	N	-,471	-,944
39	S	N	,451	N	-,451	-,907
40	S	N	,331	N	-,331	-,704
41	S	N**	,580	P	-,580	-1,175
42	S	P	,878	P	,122	,372
43	S	P	,950	P	,050	,230
44	S	P	,894	P	,106	,345
45	S	P	,726	P	,274	,614
46	S	P	,921	P	,079	,293
47	S	P	,536	P	,464	,930
48	S	P**	,200	N	,800	1,997
49	S	P**	,323	N	,677	1,449
50	S	P	,518	P	,482	,965
51	S	P	,869	P	,131	,388
52	S	P	,803	P	,197	,495
53	S	P	,942	P	,058	,248

a S = Seleccionados, N = Casos no seleccionados y ** = Casos mal clasificados.

Figura 5-16

Se observa que en el listado por casos de la Figura 5-16, para cada caso se obtiene: grupo observado, probabilidad pronosticada, grupo pronosticado, residuo y residuo tipificado.

La Figura 5-14 presenta para las variables de la ecuación su Coeficiente estimado (*B*), Error típico de *B* (*E.T.*), Estadístico de *Wald*, grados de libertad (*gl*) y p-valor para la significatividad del coeficiente estimado (*Sig*), Razón de las ventajas estimada (*exp*(*B*)) e Intervalo de confianza para *exp*(*B*) al 95%.

Según estos datos, el modelo logit estimado tiene la siguiente ecuación:

$$nodos = \frac{e^{-4,433-2,045radiografía-1,564etapa-0,761grado-0,069edad+0,024ácido}}{1+e^{-4,433-2,045radiografía-1,564etapa-0,761grado-0,069edad+0,024ácido}}$$

o la expresión equivalente:

$$nodos = \frac{1}{1+e^{-(-4,433-2,045radiografía-1,564etapa-0,761grado-0,069edad+0,024ácido)}}$$

Una vez estimado el modelo, es posible hacer predicciones sobre la probabilidad de que el cáncer se extienda a los nodos linfáticos sabiendo los valores que toman las variables independientes (edad, nivel de ácido, grado de agresividad, etapa en la que se encuentra el tumor y los resultados de una radiografía).

Por ejemplo, para un hombre de 50 años con un nivel de ácido de 40 y con valor cero en el resto de las variables, la probabilidad de que el cáncer se extienda a los nodos linfáticos es:

$$p = \frac{1}{1+e^{-(-4,433-0,069*50+0,024*40)}} = 0,000984$$

SPSS Y EL MODELO PROBIT

Se dispone del archivo *probitb.sav* que contiene datos de tres tipos de pesticida (*rotenone, deguelin* y *mixture*) que al ser aplicados en diferentes dosis sobre un número total de insectos provocan la muerte de una cierta cantidad de ellos. Mediante un modelo probit se trata de hallar la relación entre tipo de pesticida y su efecto por muerte en los insectos.

SPSS incorpora el procedimiento *Regresión Probit* que implementa el análisis de regresión probit. Este procedimiento es útil para las situaciones en las que se dispone de una variable de respuesta dicotómica que se piensa puede estar influenciada o causada por los niveles o valores de alguna o algunas variables independientes categóricas o numéricas (o de ambos tipos) y es particularmente adecuada para datos experimentales. Como estadísticos se obtienen los coeficientes de regresión y errores típicos, intersección y su error típico, *Chi-cuadrado* de Pearson de la bondad de ajuste, frecuencias observadas y esperadas e intervalos de confianza para los niveles efectivos de la variable o variables independientes. Como diagramas se obtienen los gráficos de respuestas transformadas.

Para realizar el ***análisis probit*** para nuestro ejemplo, elija en los menús *Analizar* → *Regresión* → *Probit* (Figura 5-17) y seleccione las variables y las especificaciones para el análisis (Figura 5-18). Previamente es necesario cargar en memoria el fichero de nombre *probitb.sav* mediante *Archivo* → *Abrir* → *Datos.*

En el botón *Frecuencia de respuesta* de la Figura 5-18 se introduce la variable que contiene la frecuencia de respuesta de la variable dependiente a los diferentes valores de las variables independientes, o sea, en nuestro caso la variable que contiene el número de insectos muertos al aplicar las diferentes dosis y tipos de pesticida. En el botón *Total observado* se introduce la variable que contiene el total de individuos expuestos a una combinación de niveles o valores de las variables independientes (en nuesto caso el total de insectos a los que se han aplicado los diferentes tipos y dosis de pesticida).

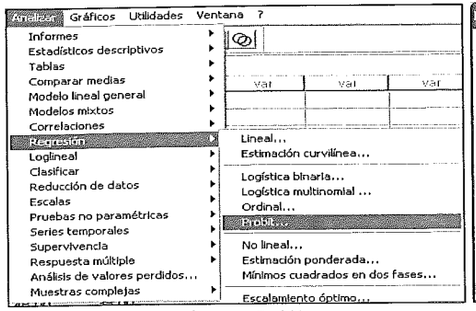


Figura 5-17

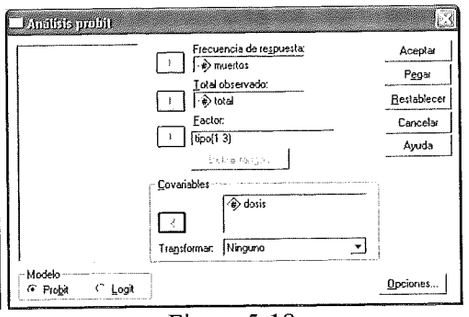


Figura 5-18

En el botón *Factor* se introduce una posible variable categórica que produce subgrupos debiendo indicarse qué niveles del factor se desea utilizar, mediante la definición del rango para el factor (botón *Definir rango*). En el campo *Covariables* introducimos otras variables independientes categóricas para el modelo. Por defecto, las covariables no se transforman. De manera opcional, puede transformar las covariables eligiendo una transformación de la lista desplegable del botón *Transformar*. En el campo *Modelo* elegiremos *Probit* para aplicar un *modelo Probit* a las proporciones de respuesta, y elegiremos *Logit* para aplicar un *modelo Logit* a las proporciones de respuesta.

El botón *Opciones* de la Figura 5-18 nos lleva a la Figura 5-19, en cuyos campos se pueden especificar opciones para el análisis probit. El campo *Estadísticos* permite solicitar los siguientes estadísticos opcionales: *Frecuencias, Potencia relativa de la mediana, Prueba de paralelismo* e *Intervalos de confianza fiduciaria*. Intervalos de confianza fiduciaria y Potencia relativa de la mediana no están disponibles si se ha seleccionado más de una covariable. Potencia relativa de la mediana y Prueba de paralelismo sólo están disponibles si se ha seleccionado una variable de factor. El campo *Tasa de respuesta natural* permite indicar una tasa de respuesta natural incluso en la ausencia del estímulo. Los posibles valores son *Ninguna, Calcular a partir de los datos* o *Valor*. El campo *Criterios* permite controlar los parámetros del algoritmo iterativo de estimación de los parámetros. Puede anular las opciones predeterminadas para *Nº máximo de iteraciones, Límite para los pasos* y *Tolerancia de la optimalidad*.

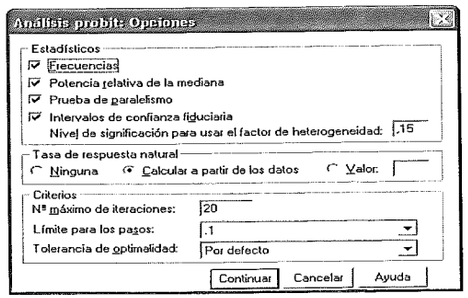


Figura 5-19

En todas las Figuras el botón *Restablecer* permite restablecer todas las opciones por defecto del sistema y elimina del cuadro de diálogo todas las asignaciones hechas con las variables. Una vez elegidas las especificaciones, se pulsa el botón *Aceptar* en la Figura 5-18 para obtener el siguiente ajuste del modelo Probit.

Estimaciones de los parámetros

Parámetro			Estimación	Error típico	Z	Sig.	Intervalo de confianza al 95%	
							Límite inferior	Límite superior
PROBIT[a]	dosis		,102	,016	6,433	,000	,071	,134
	Intersección[b]	rotenone	-,487	,311	-1,566	,117	-,798	-,176
		deguelin	-1,546	,576	-2,685	,007	-2,122	-,970
		mixture	-,745	,400	-1,862	,063	-1,145	-,345

a. Modelo PROBIT: PROBIT(p) = Intersección + BX

b. Corresponde a la variable de agrupación tipo

Según estos resultados el ajuste es bueno ya que los p-valores de los parámetros estimados son pequeños. Los modelos para cada tipo de pesticida son:

$$Probit(P_i) = -0{,}487 + 0{,}102(dosis_i)$$

$$Probit(P_i) = -1{,}546 + 0{,}102(dosis_i)$$

$$Probit(P_i) = -0{,}744 + 0{,}102(dosis_i)$$

El procedimiento también produce un gráfico que representa los valores de las probabilidades *Probit* para los diferentes valores de las variables dependientes cuantitativas (covariables), en nuestro caso la variable *dosis*, y para los diferentes valores del factor, en nuestro caso la variable *tipo* (Figura 5-20).

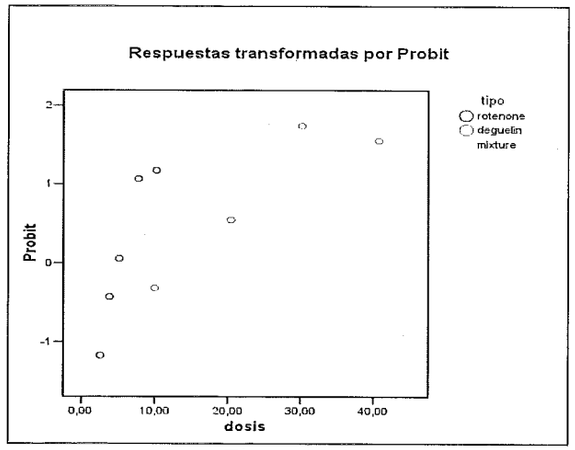


Figura 5-20

SPSS Y EL MODELO LOGIT MULTINOMIAL

Los modelos estudiados hasta ahora son modelos de elección discreta en los que el conjunto de elección tiene sólo dos alternativas posibles y que se llaman *modelos de elección binaria*. Pero cuando el conjunto de elección tiene varios valores discretos nos encontramos ante los ***modelos de elección múltiple o modelos multinomiales***.

Como ejemplo se considera el archivo *coches.sav* que contiene datos de una muestra de automóviles y se desea predecir su origen geográfico (Estados Unidos, Europa o Japón) disponiendo de datos sobre su peso, potencia y aceleración según número de cilindros.

Para realizar un análisis de regresión logística multinomial, elija en los menús *Analizar* → *Regresión* → *Logística multinomial* (Figura 5-21) y seleccione las variables y las especificaciones para el análisis (Figura 5-22). Previamente es necesario cargar en memoria el fichero de nombre *coches.sav* mediante *Archivo* → *Abrir* → *Datos*. La variable cualitativa dependiente va a ser la zona de *origen* de los automóviles, el factor va a ser el número de cilindros que tienen (*cilind*) y las variables independientes son *peso*, aceleración (*acel*) y potencia (*cv*). Se ajustará un modelo que prediga el origen de los coches por cilindros según las covariables.

En cuanto a los datos, la variable dependiente debe ser categórica. Las variables independientes pueden ser factores o covariables. En general, los factores deben ser variables categóricas y las covariables deben ser variables continuas.

En cuanto a los supuestos, se asume que la razón de ventajas de cualquier par de categorías es independiente de las demás categorías de respuesta. Según esta suposición, por ejemplo, si se introduce un nuevo producto en un mercado, la participación en el mercado de todos los demás productos quedará afectada de manera igualmente proporcional. De igual manera, dado un patrón en las covariables, se asume que las respuestas son variables multinomiales independientes.

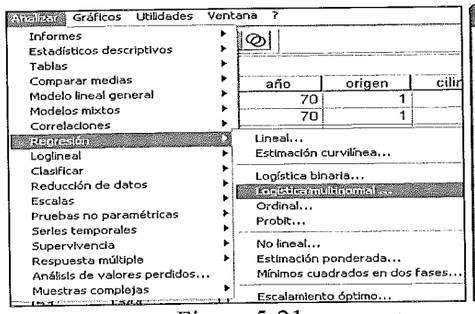


Figura 5-21

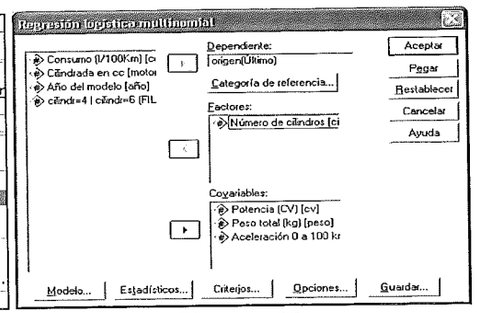


Figura 5-22

El botón *Modelos* (Figura 5-23) permite especificar diferentes modelos para una regresión logística multinomial. En *Especificar modelo* hay que tener en cuenta que un modelo de efectos principales contiene los efectos principales de las covariables y los factores, pero no contiene efectos de interacción (es la opción que vamos a elegir nosotros). Un modelo factorial completo contiene todos los efectos principales y todas las interacciones factor por factor, pero no contiene interacciones de covariable. Puede crear un modelo personalizado para especificar subconjuntos de interacciones de los factores o bien interacciones de las covariables.

Si se elige *Personalizado*, el campo *Factores y Covariables* muestra una lista de los factores y las covariables, etiquetando con (*F*) los factores fijos y con (*C*) las covariables. El campo *Modelo* sirve para diseñar el modelo, que dependerá de los efectos principales y de interacción que seleccione. La opción *Incluir la intersección en el modelo* le permite incluir o excluir del modelo un término de intersección.

El botón *Estadísticos* permite elegir distintos estadísticos de la regresión logística multinomial (Figura 5-24).

En el botón *Criterios* (Figura 5-25) puede especificar varios criterios para una regresión logística multinomial. *Iteraciones* le permite especificar el número máximo de veces que desea recorrer el algoritmo, el número máximo de pasos en la subdivisión por pasos, las tolerancias de convergencia para los cambios en el log de la verosimilitud y los parámetros y la frecuencia con que se imprime el progreso del algoritmo iterativo. *Delta* le permite especificar un valor no negativo inferior a 1. Este valor se añade a cada casilla vacía de la tabla de contingencia de las categorías de respuesta por patrones de covariables. Se ayuda así a estabilizar el algoritmo y evitar sesgos en las estimaciones. *Tolerancia para la singularidad* le permite especificar la tolerancia empleada en la comprobación de la singularidad.

El botón *Guardar* permite exportar información del modelo al archivo especificado (Figura 5-26).

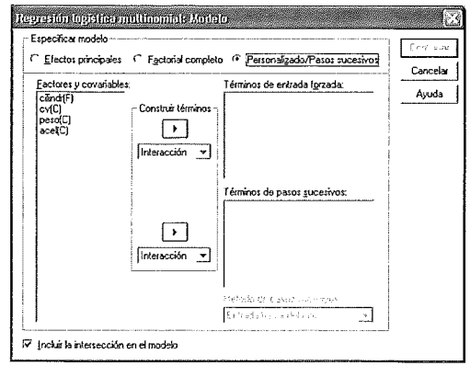


Figura 5-23

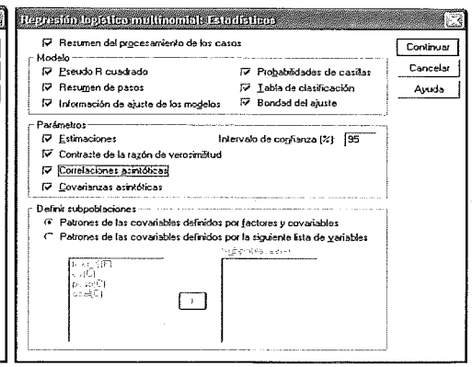


Figura 5-24

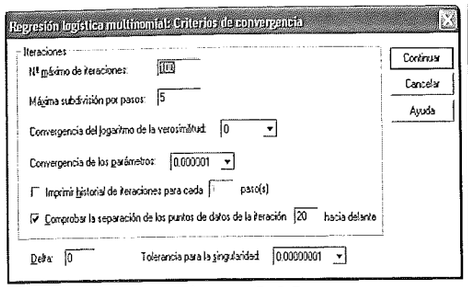


Figura 5-25

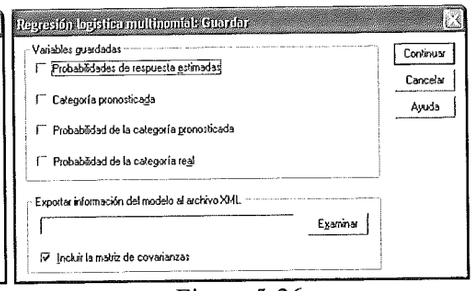


Figura 5-26

En todas las figuras el botón *Restablecer* permite restablecer todas las opciones por defecto del sistema y elimina del cuadro de diálogo todas las asignaciones hechas con las variables. Una vez elegidas las especificaciones, se pulsa el botón *Aceptar* en la Figura 5-22 para obtener los resultados del análisis según se muestra en las Figuras 5-27 a 5-31.

Se obtienen estadísticos como el historial de iteraciones (no pedido), el resumen del procesamiento de los casos (Figura 5-27), pruebas de la razón de verosimilitud para los efectos del modelo y los parciales (Figura 5-28) cuyos p-valores bajos muestran su alta significatividad, información sobre el ajuste del modelo con -2 *log* de la verosimilitud y Chi-cuadrado de la bondad de ajuste de Pearson y de la desviación que presentan p-valores correctos, así como los valores bastante aceptables de la R^2 de Cox y Snell, de Nagelkerke y de McFadden (Figura 5-29), tabla con las estimaciones de los parámetros (Figura 5-30), tabla de clasificación con frecuencias observadas respecto a las frecuencias pronosticadas por cada categoría de respuesta (Figura 5-31), tablas de contingencia para frecuencias observadas y pronosticadas (con los residuos) y proporciones por patrón en las covariables y por categoría de respuesta, matriz de covarianza asintótica y matrices de correlación (no se muestra por su elevada extensión en este caso).

Resumen del procesamiento de los casos

		N	Porcentaje marginal
País de origen	EE.UU.	249	62,4%
	Europa	71	17,8%
	Japón	79	19,8%
Número de cilindros	3 cilindros	4	1,0%
	4 cilindros	202	50,6%
	5 cilindros	3	,8%
	6 cilindros	83	20,8%
	8 cilindros	107	26,8%
Válidos		399	100,0%
Perdidos		7	
Total		406	
Subpoblación		397[a]	

a. La variable dependiente sólo tiene un valor observado en 397 (100,0%) subpoblaciones.

Figura 5-27

Contrastes de la razón de verosimilitud

Efecto	-2 log verosimilitud del modelo reducido	Chi-cuadrado	gl	Sig.
Intersección	472,104[a]	,000	0	.
cv	486,224[b]	14,120	2	,001
peso	513,231[b]	41,127	2	,000
acel	476,408[b]	4,304	2	,116
cilindr	521,631	49,527	8	,000

El estadístico de chi-cuadrado es la diferencia en las -2 log verosimilitudes entre el modelo final y el modelo reducido. El modelo reducido se forma omitiendo un efecto del modelo final. La hipótesis nula es que todos los parámetros de ese efecto son 0.

a. Este modelo reducido es equivalente al modelo final ya que la omisión del efecto no incrementa los grados de libertad.

b. Es posible que haya una separación casi completa de los datos. O bien no existen estimaciones de máxima verosimilitud o bien algunas estimaciones de parámetros son infinitas.

Figura 5-28

Información del ajuste del modelo

Modelo	-2 log verosimilitud	Chi-cuadrado	gl	Sig.
Sólo la intersección	735,826			
Final	472,104	263,723	14	,000

Bondad de ajuste

	Chi-cuadrado	gl	Sig.
Pearson	515,743	778	1,000
Desviación	472,104	778	1,000

Pseudo R-cuadrado

Cox y Snell	,484
Nagelkerke	,575
McFadden	,358

Figura 5-29

Estimaciones de los parámetros

País de origen(a)		B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
								Límite inferior	Límite superior
EE.UU.	Intersección	19,355	3224,807	,000	1	,995			
	cv	-,074	,020	13,294	1	,000	,929	,893	,966
	peso	,015	,003	31,042	1	,000	1,015	1,010	1,021
	acel	-,196	,102	3,651	1	,056	,822	,673	1,005
	[cilindr=3]	-40,898	7574,598	,000	1	,996	1,731E-18	,000	.(b)
	[cilindr=4]	-21,976	3224,806	,000	1	,995	2,858E-10	,000	.(b)
	[cilindr=5]	-21,918	39446,04	,000	1	1,000	3,027E-10	,000	.(b)
	[cilindr=6]	-21,566	3224,806	,000	1	,995	4,306E-10	,000	.(b)
	[cilindr=8]	0(c)	.	.	0	.	.	.	.
Europa	Intersección	-4,176	2,575	2,631	1	,105			
	cv	-,044	,021	4,178	1	,041	,957	,918	,998
	peso	,011	,003	15,872	1	,000	1,011	1,006	1,017
	acel	-,045	,102	,198	1	,657	,956	,783	1,167
	[cilindr=3]	-20,400	,000	.	1	.	1,381E-09	1,381E-09	1,381E-09
	[cilindr=4]	-,255	,835	,093	1	,760	,775	,151	3,982
	[cilindr=5]	19,962	38148,025	,000	1	1,000	467109975,38	,000	.(b)
	[cilindr=6]	-2,216	,000	.	1	.	,109	,109	,109
	[cilindr=8]	0(c)	.	.	0	.	.	.	.

a La categoría de referencia es: Japón.
b Se ha producido un desbordamiento de punto flotante al calcular este estadístico. Por lo tanto, el valor asignado ha sido el valor perdido del sistema.
c Este parámetro se ha establecido a cero porque es redundante.

Figura 5-30

Clasificación

	Pronosticado			
Observado	EE.UU.	Europa	Japón	Porcentaje correcto
EE.UU.	219	8	22	88,0%
Europa	27	15	29	21,1%
Japón	23	3	53	67,1%
Porcentaje global	67,4%	6,5%	26,1%	71,9%

Figura 5-31

El cuadro de estimaciones de los parámetros de la Figura 5-30 muestra para cada variable de la ecuación el coeficiente estimado (*B*), el error típico de *B*, el estadístico de Wald, los grados de libertad, los p-valores que miden el nivel de significación de los parámetros estimados del modelo (*Sig.*), la razón de las ventajas estimada (*exp*(*B*)) e Intervalo de confianza al 95% para *exp*(*B*). Se observa la fuerte significatividad de las covariables y la baja sisgnificatividad de la constante y de los niveles del factor. Ello puede inducirnos a cambiar el factor de clasificación o a no utilizarlo.

La tabla de clasificación de la Figura 5-31 muestra que el poder clasificativo de nuestro modelo logit multinomial es bastante bueno, ya que ha clasificado de modo correcto un 72% de las observaciones conocidas, esperándose que ese poder clasificativo se proyecte a estimaciones futuras realizadas con el modelo.

Como alternativa podemos estimar el modelo sin variable factor rellenando la pantalla de entrada del procedimiento como se indica en la Figura 5-32.

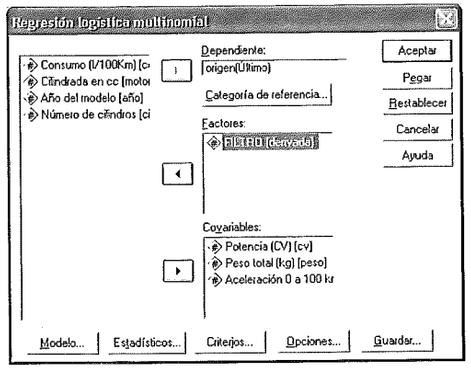


Figura 5-32

Al pulsar *Aceptar* se observa que la significatividad de los parámetros estimados ha aumentado mucho (Figura 5-33).

Estimaciones de los parámetros

País de origen[a]		B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B) Límite inferior	Límite superior
EE.UU.	Intersección	-5,667	1,991	8,099	1	,004			
	cv	-,041	,015	7,144	1	,008	,960	,932	,989
	peso	,014	,002	46,344	1	,000	1,015	1,010	1,019
	acel	-,135	,096	1,987	1	,159	,874	,724	1,054
Europa	Intersección	-3,875	2,059	3,543	1	,060			
	cv	-,037	,018	4,185	1	,041	,964	,930	,998
	peso	,008	,002	12,708	1	,000	1,008	1,004	1,013
	acel	,028	,096	,083	1	,773	1,028	,852	1,240

a. La categoría de referencia es: Japón.

Figura 5-33

En las estimaciones anteriores la categoría de referencia es Japón. Para tomar como categoría de referencia Estados Unidos. (valor 1 de la variable origen), rellenamos la pantalla de entrada del procedimiento como se indica en la Figura 5-34 (obsérvese la cumplimentación del botón *Categoría de referencia*) y al pulsar *Aceptar*, se obtienen las estimaciones de la Figura 5-35. Para tomar como categoría de referencia Europa (valor 2 de la variable origen), rellenamos la pantalla de entrada del procedimiento como se indica en la Figura 5-36 y al pulsar *Aceptar*, se obtienen las estimaciones de la Figura 5-37. Se observa que las mejores significatividades para los parámetros del modelo se obtienen cuando la categoría de referencia es Japón (Figura 5-33). En este caso sólo la variable aceleración para los automóviles europeos es no significativa, aunque para los automóviles de Estados unidos también es esta variable la menos significativa (85%).

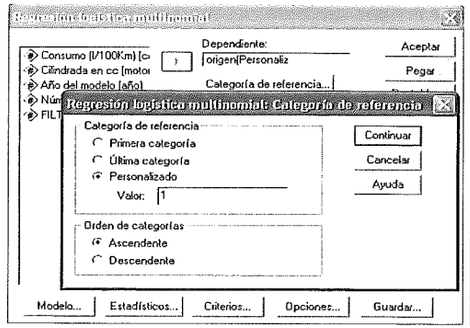


Figura 5-34

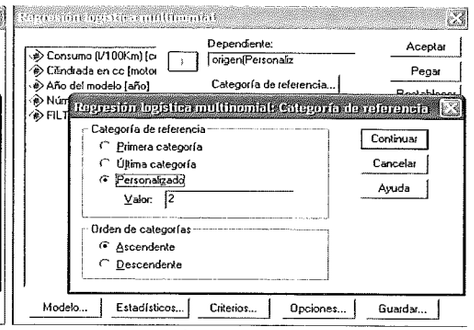


Figura 5-35

Estimaciones de los parámetros

País de origen[a]		B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B) Límite inferior	Límite superior
Europa	Intersección	1,791	1,861	,927	1	,336			
	cv	,004	,016	,055	1	,814	1,004	,973	1,036
	peso	-,006	,002	14,997	1	,000	,994	,990	,997
	acel	,163	,093	3,052	1	,081	1,177	,980	1,412
Japón	Intersección	5,667	1,991	8,099	1	,004			
	cv	,041	,015	7,144	1	,008	1,041	1,011	1,073
	peso	-,014	,002	46,344	1	,000	,986	,982	,990
	acel	,135	,096	1,987	1	,159	1,145	,949	1,381

a. La categoría de referencia es: EE.UU..

Figura 5-36

Estimaciones de los parámetros

País de origen[a]		B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B) Límite inferior	Límite superior
EE.UU.	Intersección	-1,791	1,861	,927	1	,336			
	cv	-,004	,016	,055	1	,814	,996	,966	1,028
	peso	,006	,002	14,997	1	,000	1,006	1,003	1,010
	acel	-,163	,093	3,052	1	,081	,850	,708	1,020
Japón	Intersección	3,875	2,059	3,543	1	,060			
	cv	,037	,018	4,185	1	,041	1,038	1,002	1,075
	peso	-,008	,002	12,708	1	,000	,992	,987	,996
	acel	-,028	,096	,083	1	,773	,973	,807	1,173

a. La categoría de referencia es: Europa.

Figura 5-37

La ecuación del mejor modelo Logit multinomial estimado (Figura 5-33) es:

$$EE.UU. = \frac{1}{1+e^{-(-5,667-0,41cv+0,14peso-0,135acel}}, \quad Europa = \frac{1}{1+e^{-(-3,875-0,37cv+0,08peso+0,28acel}}$$

Por tanto, para hallar la probabilidad de que un automóvil proceda de Estados Unidos, dados los valores de su potencia, peso y aceleración, se sustituyen esos valores en la primera ecuación. Si se quiere la probabilidad de procedencia de Europa se utiliza la segunda ecuación.

SAS Y LA REGRESIÓN LOGÍSTICA. PROC LOGISTIC

El PROC LOGISTIC ajusta el modelo logístico. Su sintaxis es la siguiente:

PROC LOGISTIC opciones;
BY variables;
CLASS variable (v-opciones) variable (v-opciones)... / v-opciones;
CONTRAST 'etiqueta' effect values ,... effect values /opciones;
FREQ variable;
MODEL response = effects / opciones;
MODEL events/trials = effects / opciones;
OUTPUT OUT=SAS-data-set estadísticos=nombres / opción;
label: TEST equation1 , ... , equationk /opción;
UNITS independent1 = list1 ... independentk = listk /opción;
WEIGHT variable / opción;

Las **opciones de PROC LOGISTIC** son DATA=conjunto de datos de entrada, OUTEST=conjunto de datos que contiene los parámetros estimados y otros estadísticos del ajuste, INEST=conjunto de datos que contiene las estimaciones iniciales para los parámetros del modelo, COVOUT (guarda la matriz de covarianzas estimada, SIMPLE (muestra estadísticos descriptivos simples), NOPRINT (elimina la salida), DESCENDING (cambia el orden para los niveles de la variable respuesta), NAMELEN=longitud de los nombres de los efectos en las tablas, ORDER=DATA | FORMATTED | FREQ | INTERNAL, ALPHA=nivel de significación y EXACTONLY (sólo estimaciones exactas).

La **sentencia CLASS** especifica las variables de clasificación a utilizar en el análisis. Sus opciones más importantes son PARAM=EFFECT|GLM|POLY| ORTHPOLY|REF (métodos de parametrización para la clasificación de las variables), REF = 'nivel'| estadístico (nivel de referencia para EFFECT o REFF), DESCENDING (cambia el orden para los niveles de las variables de clasificación) y ORDER=DATA | FORMATTED | FREQ | INTERNAL.

La sentencia **CONTRAST** permite realizar contrastes de los parámetros del modelo. Sus opciones son ALPHA=nivel de significación, E para presentar la matriz L, ESTIMATE=PARM|EST|BOTH (contraste normal o exponencial) y SINGULAR=valor entre 0 y 1 (controla las singularidades).

La sentencia **MODEL** especifica la variable respuesta y los efectos explicativos, incluyendo covariantes, efectos principales, interacciones y efectos anidados. Sus opciones son las siguientes:

Opciones de especificación del modelo

LINK= CLOGLOG | LOGIT | PROBIT (función de enlace)
NOINT: suprime la constante del modelo
NOFIT: suprime el ajuste del modelo
OFFSET= variable *offset*
SELECTION= BACKWARD | B (método de selección de variable)

Opciones de selección de variables

BEST= nº de modelos mostrados por la selección SCORE
DETAILS: requiere resultados detallados en cada paso
FAST: utiliza el método de eliminación rápida
HIERARCHY= NONE | SINGLE | SINGLECLASS | MULTIPLE |
MULTIPLECLASS (especifica si se usa modelo jerárquico, de efectos simples o de efectos múltiples).
INCLUDE= nº de primeros efectos a incluir en el modelo
MAXSTEP= máximo número de pasos para la selección STEPWISE
SEQUENTIAL añade o borra variables en orden secuencial
SLENTRY= nivel de significación para variables entrantes
SLSTAY= nivel de significación para variables salientes
START= nº de variables en el primer modelo
STOP= nº de variables en el modelo final
STOPRES añade o borra variables por el criterio de la Chi-cuadrado de los residuos

Opciones de especificación del ajuste del modelo

ABSFCONV= nº que especifica el criterio de convergencia (función absoluta)
FCONV= nº que especifica el criterio de convergencia (función relativa)
GCONV= nº que especifica el criterio de convergencia (gradiente relativo)
XCONV= nº que especifica el criterio de convergencia (parámetro relativo)
MAXITER= número máximo de iteraciones
NOCHECK: suprime chequeo para parámetro infinito
RIDGING= ABSOLUTE | RELATIVE | NONE (máxima verosimilitud)
SINGULAR= tolerancia para contraste de singularidad
TECHNIQUE= FISHER | NEWTON (algoritmo iterativo para maximización)

Opciones de intervalos de confianza

ALPHA= nivel de significación
CLPARM= PL | WALD | BOTH (intervalos de confianza para los parámetros)
CLODDS= PL | WALD | BOTH (intervalos de confianza para *odds ratios*)
PLCONV= nº que da el criterio para el perfil de convergencia de verosimilitud

Opciones para clasificación de observaciones

CTABLE: muestra la tabla de clasificación
PEVENT= especifica probabilidad del evento principal
PPROB= especifica probabilidad para clasificación de puntos de corte

Opciones para el contraste de la bondad de ajuste

AGGREGATE: determina subpoblaciones para la Chi-cuadrado de Pearson
SCALE= DEVIANCE o D | PEARSON o P | WILLIAMS | NONE o N | constante (especifica método para corregir la sobredispersión)
LACKFIT: requiere el contraste de la bondad de ajuste de Hosmer y Lemeshow

Otras opciones

ROCEPS= criterio para agrupar probabilidades de eventos
INFLUENCE: muestra estadísticos de influencia
IPLOTS: requiere gráficos de índices
CORRB: muestra matriz de correlaciones
COVB: muestra matriz de covarianzas
EXPB: muestra valores exponenciales estimados
ITPRINT: muestra historial de iteraciones
NODUMMYPRINT: suprime la tabla de información del nivel de clase
PARMLABEL: muestra etiquetas de parámetro
RSQUARE: muestra R^2 generalizado
STB: muestra estimaciones estándar

La **opción OUTPUT** permite especificar el conjunto de datos en el que se guardarán resultados del ajuste. Estos resultados son los siguientes:

Estadísticos cuando la respuesta es binaria u ordinal : LOWER (límite inferior de confianza para la probabilidad), PREDICTED (probabilidad predicha de un evento), PREDPROBS (individuales o I, acumulativas o C y cruzadas o X), STDXBETA (error estándar estimado de XBETA), UPPER (límite superior de confianza para la probabilidad), XBETA (estimación del predictor lineal).

Estadísticos cuando la respuesta es sólo binaria: C (intervalos de confianza de influencia en las estimaciones de las observaciones individuales), CBAR (otros intervalos de confianza de diagnóstico), DFBETAS (diferencias estandarizadas), DIFCHISQ (cambio en la Chi-cuadrado de la bondad de ajuste), DIFDEV (cambio de la devianza), H (elemento de la diagonal de la matriz para detectar puntos extremos), RESCHI (residuo Chi-cuadrado de Pearson para identificar observaciones mal introducidas en el modelo), RESDEV (devianza residual para identificar observaciones mal introducidas en el modelo).

OTRAS opciones: ALPHA=nivel de significación.

La **sentencia TEST** se utiliza para especificar los contrastes a realizar con los parámetros de la regresión. Por ejemplo, si el modelo es $y = a1\ a2\ a3\ a4$, serían restricciones válidas: *test1: test intercept + .5 * a2 = 0*, o *test2: test intercept + .5 * a2*, o *test3: test a1=a2=a3* o *test4: test a1=a2, a2=a3*, o cualquier otra similar. La opción PRINT de esta sentencia muestra determinados cálculos intermedios.

La **sentencia UNITS** permite especificar unidades de cambio para variables continuas explicativas tales como *odds ratios* a medida que pueden ser estimados. La opción DEFAULT permite introducir una lista de unidades de cambio para las variables explicativas no especificadas en UNITS.

La **sentencia WEIGHT** permite introducir pesos para las observaciones. La opción NORMALIZE indica pesos normalizados.

Como ejemplo consideramos datos de un estudio de remisión del cáncer en varios pacientes en los que se han medido las variables *remisión* (1 si remite y 0 si no remite) y varios factores de riesgo (*f1* a *f6*). A partir de estos datos ajustaremos un modelo logístico que nos dé la probabilidad de remisión del cáncer en un paciente con factores de riesgo dados. El ajuste se hará por regresión paso a paso y se obtendrán matrices de covarianzas, probabilidades predichas e intervalos de confianza.

```
data cancer;
      input remision f1-f6;
      label remision='Remision completa';
      datalines;
  1   .8   .83  .66  1.9  1.1     .996
  1   .9   .36  .32  1.4   .74    .992
  0   .8   .88  .7    .8   .176   .982
  0  1     .87  .87   .7  1.053   .986
  1   .9   .75  .68  1.3   .519   .98
  0  1     .65  .65   .6   .519   .982
  1   .95  .97  .92  1    1.23    .992
  0   .95  .87  .83  1.9  1.354  1.02
  0  1     .45  .45   .8   .322   .999
  0   .95  .36  .34   .5  0      1.038
  0   .85  .39  .33   .7   .279   .988
  0   .7   .76  .53  1.2   .146   .982
  0   .8   .46  .37   .4   .38   1.006
  0   .2   .39  .08   .8   .114   .99
  0  1     .9   .9   1.1  1.037   .99
  1  1     .84  .84  1.9  2.064  1.02
  0   .65  .42  .27   .5   .114  1.014
  0  1     .75  .75  1    1.322  1.004
  0   .5   .44  .22   .6   .114   .99
  1  1     .63  .63  1.1  1.072   .986
  0  1     .33  .33   .4   .176  1.01
  0   .9   .93  .84   .6  1.591  1.02
  1  1     .58  .58  1     .531  1.002
  0   .95  .32  .3   1.6   .886   .988
  1  1     .6   .6   1.7   .964   .99
  1  1     .69  .69   .9   .398   .986
  0  1     .73  .73   .7   .398   .986
 ;
```

```
title 'Regresion paso a paso';
   proc logistic data=cancer descending outest=betas covout;
      model remision=f1-f6
                  / selection=stepwise
                    slentry=0.3
                    slstay=0.35
                    details
                    lackfit;
      output out=pred p=phat lower=lcl upper=ucl
             predprobs=(individual crossvalidate);
   run;
   proc print data=betas;
      title2 'Parametros estimados y matriz de covarianzas';
   run;
   proc print data=pred;
     title2 'Probabilidades predichas e intervalos de confianza al 95%';
   run;
```

Después de 4 pasos para la convergencia, la salida es la siguiente:

Regresion paso a paso

The LOGISTIC Procedure
Model Information

Data Set	WORK.CANCER	
Response Variable	remision	Remision completa
Number of Response Levels	2	
Number of Observations	27	
Link Function	Logit	
Optimization Technique	Fisher's scoring	

Response Profile

Ordered Value	remision	Total Frequency
1	1	9
2	0	18

Stepwise Selection Procedure

Step 3. Effect f1 entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.
Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	36.372	29.953
SC	37.668	35.137
-2 Log L	34.372	21.953

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	12.4184	3	0.0061
Score	9.2502	3	0.0261
Wald	4.8281	3	0.1848

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	67.6339	56.8875	1.4135	0.2345
f1	1	9.6521	7.7511	1.5507	0.2130
f4	1	3.8671	1.7783	4.7290	0.0297
f6	1	-82.0737	61.7124	1.7687	0.1835

Regresion paso a paso
The LOGISTIC Procedure
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
f1	>999.999	0.004	>999.999
f4	47.804	1.465	>999.999
f6	<0.001	<0.001	>999.999

Association of Predicted Probabilities and Observed Responses

Percent Concordant	88.9	Somers' D	0.778
Percent Discordant	11.1	Gamma	0.778
Percent Tied	0.0	Tau-a	0.359
Pairs	162	c	0.889

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
0.1831	3	0.9803

Analysis of Effects in Model

Effect	DF	Wald Chi-Square	Pr > ChiSq
f1	1	1.5507	0.2130
f4	1	4.7290	0.0297
f6	1	1.7687	0.1835

Analysis of Effects Not in the Model

Effect	DF	Score Chi-Square	Pr > ChiSq
f2	1	0.0956	0.7572
f3	1	0.0844	0.7714
f5	1	0.0208	0.8852

NOTE: No (additional) effects met the 0.3 significance level for entry into the model.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	f4		1	1	7.9311	.	0.0049
2	f6		1	2	1.2591	.	0.2618
3	f1		1	3	1.4700	.	0.2254

Partition for the Hosmer and Lemeshow Test

Group	Total	remision = 1 Observed	remision = 1 Expected	remision = 0 Observed	remision = 0 Expected
1	4	0	0.00	4	4.00
2	3	0	0.03	3	2.97
3	3	0	0.34	3	2.66
4	3	1	0.65	2	2.35
5	3	0	0.84	3	2.16
6	3	2	1.35	1	1.65
7	3	2	1.84	1	1.16
8	3	3	2.15	0	0.85
9	2	1	1.80	1	0.20

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
7.1966	7	0.4087

Parametros estimados y matriz de covarianzas

Obs	_LINK_	_TYPE_	_STATUS_	_NAME_	Intercept	f1	f2	f3	f4	f5	f6	_LNLIKE_
1	LOGIT	PARMS	0 Converged	remision	67.63	9.652	.	.	3.8671	.	-82.07	-10.9767
2	LOGIT	COV	0 Converged	Intercept	3236.19	157.097	.	.	64.5726	.	-3483.23	-10.9767
3	LOGIT	COV	0 Converged	f1	157.10	60.079	.	.	6.9454	.	-223.67	-10.9767
4	LOGIT	COV	0 Converged	f2	.	.	.	.	.	.	.	-10.9767
5	LOGIT	COV	0 Converged	f3	.	.	.	.	.	.	.	-10.9767
6	LOGIT	COV	0 Converged	f4	64.57	6.945	.	.	3.1623	.	-75.35	-10.9767
7	LOGIT	COV	0 Converged	f5	.	.	.	.	.	.	.	-10.9767
8	LOGIT	COV	0 Converged	f6	-3483.23	-223.669	.	.	-75.3513	.	3808.42	-10.9767

Probabilidades predichas e intervalos de confianza al 95%

Obs	remision	f1	f2	f3	f4	f5	f6	_FROM_	_INTO_	IP_1	IP_0	XP_1	XP_0	_LEVEL_	phat	lcl	ucl
1	1	0.80	0.83	0.66	1.9	1.100	0.996	1	1	0.72265	0.27735	0.56127	0.43873	1	0.72265	0.16892	0.97093
2	1	0.90	0.36	0.32	1.4	0.740	0.992	1	1	0.57874	0.42126	0.52539	0.47461	1	0.57874	0.26788	0.83762
3	0	0.80	0.88	0.70	0.8	0.176	0.982	0	0	0.10460	0.89540	0.12940	0.87060	1	0.10460	0.00781	0.63419
4	0	1.00	0.87	0.87	0.7	1.053	0.986	0	0	0.28258	0.71742	0.32741	0.67259	1	0.28258	0.07498	0.65683
5	1	0.90	0.75	0.68	1.3	0.519	0.980	1	1	0.71418	0.28582	0.63099	0.36901	1	0.71418	0.25218	0.94876
6	0	1.00	0.65	0.65	0.6	0.519	0.982	0	0	0.27089	0.72911	0.32731	0.67269	1	0.27089	0.05852	0.68951
7	1	0.95	0.97	0.92	1.0	1.230	0.992	1	0	0.32156	0.67844	0.27077	0.72923	1	0.32156	0.13255	0.59516
8	0	0.95	0.87	0.83	1.9	1.354	1.020	0	1	0.60723	0.39277	0.90094	0.09906	1	0.60723	0.10572	0.95287
9	0	1.00	0.45	0.45	0.8	0.322	0.999	0	0	0.16632	0.83368	0.19136	0.80864	1	0.16632	0.03018	0.56123
10	0	0.95	0.36	0.34	0.5	0.000	1.038	0	0	0.00157	0.99843	0.00160	0.99840	1	0.00157	0.00000	0.68962
11	0	0.85	0.39	0.33	0.7	0.279	0.988	0	0	0.07285	0.92715	0.08277	0.91723	1	0.07285	0.00614	0.49982
12	0	0.70	0.76	0.53	1.2	0.146	0.982	0	0	0.17286	0.82714	0.36162	0.63838	1	0.17286	0.00637	0.87206
13	0	0.80	0.46	0.37	0.4	0.380	1.006	0	0	0.00346	0.99654	0.00356	0.99644	1	0.00346	0.00001	0.46530
14	0	0.20	0.39	0.08	0.8	0.114	0.990	0	0	0.00018	0.99982	0.00019	0.99981	1	0.00018	0.00000	0.96482
15	0	1.00	0.90	0.90	1.1	1.037	0.990	0	1	0.57122	0.42878	0.64646	0.35354	1	0.57122	0.25303	0.83973
16	1	1.00	0.84	0.84	1.9	2.064	1.020	1	1	0.71470	0.28530	0.52787	0.47213	1	0.71470	0.15362	0.97189
17	0	0.65	0.42	0.27	0.5	0.114	1.014	0	0	0.00062	0.99938	0.00063	0.99937	1	0.00062	0.00000	0.62665
18	0	1.00	0.75	0.75	1.0	1.322	1.004	0	0	0.22289	0.77711	0.26388	0.73612	1	0.22289	0.04483	0.63670
19	0	0.50	0.44	0.22	0.6	0.114	0.990	0	0	0.00154	0.99846	0.00158	0.99842	1	0.00154	0.00000	0.79644
20	1	1.00	0.63	0.63	1.1	1.072	0.986	1	1	0.64911	0.35089	0.57947	0.42053	1	0.64911	0.26305	0.90555
21	0	1.00	0.33	0.33	0.4	0.176	1.010	0	0	0.01693	0.98307	0.01830	0.98170	1	0.01693	0.00029	0.50475
22	0	0.90	0.93	0.84	0.6	1.591	1.020	0	0	0.00622	0.99378	0.00652	0.99348	1	0.00622	0.00003	0.56062
23	1	1.00	0.58	0.58	1.0	0.531	1.002	1	0	0.25261	0.74739	0.15577	0.84423	1	0.25261	0.06137	0.63597
24	0	0.95	0.32	0.30	1.6	0.886	0.988	0	1	0.87011	0.12989	0.96363	0.03637	1	0.87011	0.40910	0.98481
25	1	1.00	0.60	0.60	1.7	0.964	0.990	1	1	0.93132	0.06868	0.91983	0.08017	1	0.93132	0.44114	0.99573
26	1	1.00	0.69	0.69	0.9	0.398	0.986	1	0	0.46051	0.53949	0.37688	0.62312	1	0.46051	0.16612	0.78529
27	0	1.00	0.73	0.73	0.7	0.398	0.986	0	0	0.28258	0.71742	0.32741	0.67259	1	0.28258	0.07498	0.65683

Se observa que las variables que se mantienen en el modelo para obtener significatividad de sus parámetros al 70% son: la constante, *f1, f4 y f6* (p-valores menores que 0,3 en la estimación por máxima verosimilitud en el paso 3). Asímismo, el modelo es significativo globalmente al 80% (p-valores en los contrastes de razón de verosimilitud, *score* y Wald menores que 0,2 en el paso 3). El modelo final estimado que nos de la probabilidad de remisión del cáncer en un paciente con factores de riesgo dados *f1, f4* y *f6* es el siguiente:

$$p = \frac{1}{1+e^{-(67,73+9,65f1+3,86f4-82,07f6)}}$$

SAS Y EL MODELO PROBIT: PROCEDIMIENTO PROBIT

El procedimiento PROBIT permite la estimación de probabilidades mediante estimación por máxima verosimilitud en un modelo lineal general con función de distribución normal. Su sintaxis es la siguiente:

PROC PROBIT opciones;
CLASS variables;
MODEL respuesta=independientes / opciones ;
BY variables;
OUTPUT OUT=conjunto de datos de salida / estadísticos=nombres ;
WEIGHT variable

Las **opciones de PROC PROBIT** son: DATA=conjunto de datos de entrada, OUTEST=conjunto de datos que contiene los parámetros estimados y otros estadísticos del ajuste, INVERSECL (halla límites de confianza para los valores de las primeras variables continuas independientes que producen tasas de respuesta seleccionadas), COVOUT (guarda la matriz de covarianzas estimada, C=tasa de respuesta natural, OPTC (se estima automáticamente la tasa de respuesta natural), HPROB=mínimo nivel de significación para el contraste de la Chi-cuadrado de la bondad del ajuste, LOG (reemplaza las primeras variables independientes continuas por sus logaritmos naturales), LOG10 (especifica que los análisis se hagan en logaritmos decimales), NOPRINT (elimina la salida) y ORDER=DATA | FORMATTED | FREQ | INTERNAL).

Las **opciones de la sentencia MODEL** son: CONVERGE = valor de cambio entre dos estimaciones consecutivas que hace que el algoritmo de Newton Rhampson finalice, CORRB (muestra la matriz de correlaciones de los parámetros estimados), COVB (muestra la matriz de covarianzas de los parámetros estimados), DISTRIBUTION=NORMAL | LOGISTIC | EXTREME o EXTREMEVALUE o GOMPERTZ indica la función de distribución utilizada en el modelo de respuesta lineal de probabilidad, HPROB=mínimo nivel de significación para el contraste de la Chi-cuadrado de la bondad del ajuste, INITIAL=lista de valores iniciales para los parámetros, INTERCEP=valor (inicializa el parámetro constante), INVERSECL (halla límites de confianza para los valores de las primeras variables continuas independientes que producen tasas de respuesta seleccionadas), ITPRINT (muestra la historia de iteraciones), LACKFIT (realiza dos test de la bondad de ajuste de la Chi-cuadrado), MAXITER=máximo número de iteraciones, NOINT (ajusta el modelo sin constante y SINGULAR=valor del criterio de singularidad.

Los **estadísticos a utilizar OUTPUT** son: PROB | P (probabilidades acumuladas estimadas $Pi=c+(1-c)F(x_i\beta)$ que es la forma que toma el modelo Probit cuando la variable respuesta es binaria siendo c la tasa natural de respuesta), STD (error estándar de las estimaciones $x_i\beta$) y XBETA (estimaciones de $x_i\beta$).

Como ejemplo consideramos el mismo que para el modelo logit. El ajuste a un modelo probit se hará de la siguiente forma:

```
data cancer;
     input remision f1-f6;
     label remision='Remision completa';
     datalines;
  1   .8    .83   .66   1.9   1.1      .996
  1   .9    .36   .32   1.4    .74     .992
  0   .8    .88   .7     .8    .176    .982
  0  1      .87   .87    .7   1.053    .986
  1   .9    .75   .68   1.3    .519    .98
  0  1      .65   .65    .6    .519    .982
  1   .95   .97   .92   1     1.23     .992
  0   .95   .87   .83   1.9   1.354   1.02
  0  1      .45   .45    .8    .322    .999
  0   .95   .36   .34    .5   0       1.038
  0   .85   .39   .33    .7    .279    .988
  0   .7    .76   .53   1.2    .146    .982
  0   .8    .46   .37    .4    .38    1.006
  0   .2    .39   .08    .8    .114    .99
  0  1      .9    .9    1.1   1.037    .99
  1  1      .84   .84   1.9   2.064   1.02
  0   .65   .42   .27    .5    .114   1.014
  0  1      .75   .75   1     1.322   1.004
  0   .5    .44   .22    .6    .114    .99
  1  1      .63   .63   1.1   1.072    .986
  0  1      .33   .33    .4    .176   1.01
  0   .9    .93   .84    .6   1.591   1.02
  1  1      .58   .58   1      .531   1.002
  0   .95   .32   .3    1.6    .886    .988
  1  1      .6    .6    1.7    .964    .99
  1  1      .69   .69    .9    .398    .986
  0  1      .73   .73    .7    .398    .986
 ;
title 'Ajuste Probit';
   proc probit data=cancer;
      model remision=f1-f6 / lackfit;
   run;
```

La salida es la siguiente:

```
                    Procedimiento Probit

                  Información del modelo

Conjunto de datos              WORK.CANCER
Variable dependiente             remision    Remision completa
Número de observaciones                27
Nombre de la distribución           Normal
Verosimilitud log             -10.82921779

            Número de observaciones leídas          27
            Número de observaciones usadas          27
```



```
                  Información de nivel de clase

              Nombre        Niveles    Valores

              remision            2    0 1

                     Perfil de respuesta

                Valor                  Frecuencia
             ordenado    remision          total

                    1    0                    18
                    2    1                     9

PROC PROBIT está modelando las probabilidades de niveles de remision que tienen valores ordenados
LOWER en la tabla de perfil de respuesta.

        Algoritmo convergido.

                     Tipo III Análisis de efectos

                                Chi-cuadrado
                Efecto      DF        de Wald     Pr > ChiSq

                f1           1         0.2384        0.6254
                f2           1         0.0893        0.7651
                f3           1         0.0776        0.7805
                f4           1         2.8226        0.0929
                f5           1         0.0003        0.9867
                f6           1         1.6441        0.1998

        Análisis del estimador del parámetro de máxima verosimilitud

                                  Error    95% Límites        Chi-
        Parámetro  DF Estimador estándar    de confianza   cuadrado Pr > ChiSq

        Intercept   1  -32.8402  41.8239 -114.814  49.1331     0.62     0.4323
        f1          1  -13.7495  28.1612 -68.9445  41.4454     0.24     0.6254
        f2          1  -10.2893  34.4304 -77.7718  57.1931     0.09     0.7651
        f3          1   10.1819  36.5423 -61.4397  81.8035     0.08     0.7805
        f4          1   -2.3532   1.4007  -5.0984   0.3921     2.82     0.0929
        f5          1   -0.0228   1.3667  -2.7016   2.6559     0.00     0.9867
        f6          1   49.5914  38.6766 -26.2134 125.3962     1.64     0.1998
```

Se observa que las únicas variables significativas son f1 y f6. Podríamos intentar reestimar el modelo sólo con estas variables añadiendo como mucho la constante y f1 (por si hay cambios en la significatividad al eliminar variables). Tendríamos la siguiente sintaxis:

```
proc probit data=cancer;
     model remision=f1 f4 f6 / lackfit;
  run;
```

La salida es la siguiente:

```
                    Procedimiento Probit
                  Información del modelo

Conjunto de datos                WORK.CANCER
Variable dependiente               remision     Remision completa
Número de observaciones                  27
Nombre de la distribución            Normal
Verosimilitud log              -10.95007639

          Número de observaciones leídas          27
          Número de observaciones usadas          27

              Información de nivel de clase

              Nombre        Niveles    Valores

              remision            2    0 1

                   Perfil de respuesta
                Valor                  Frecuencia
              ordenado    remision        total

                     1    0                    18
                     2    1                     9

PROC PROBIT está modelando las probabilidades de niveles de remision que tienen valores ordenados
LOWER en la tabla de perfil de respuesta.

     Algoritmo convergido.

                  Tipo III Análisis de efectos

                                Chi-cuadrado
          Efecto        DF          de Wald     Pr > ChiSq

          f1             1           1.4736         0.2248
          f4             1           5.2886         0.0215
          f6             1           1.6751         0.1956

     Análisis del estimador del parámetro de máxima verosimilitud

                                Error    95% Límites        Chi-
  Parámetro  DF Estimador estándar    de confianza    cuadrado Pr > ChiSq

  Intercept   1  -36.7548  32.3607 -100.181  26.6711       1.29     0.2560
  f1          1   -5.6298   4.6376 -14.7194   3.4598       1.47     0.2248
  f4          1   -2.2513   0.9790  -4.1700  -0.3326       5.29     0.0215
  f6          1   45.1815  34.9095 -23.2398 113.6029       1.68     0.1956
```

Se observa que las variables significativas son las mismas que en el modelo logit y con p-valores semejantes, lo que indica que es indiferente utilizar un modelo logit o un modelo probit. La ecuación del modelo probit estimado será la siguiente:

$$p = Pr(\,Remisión) = F\,(\,-36{,}75 - 5{,}62\,f1 - 2{,}25\,f4 + 45{,}18\,f6) =$$

$$= (2\pi)^{-1/2} \int_{-\infty}^{-36{,}75 - 5{,}62\,f1 - 2{,}25\,f4 + 45{,}18\,f6} e^{-\frac{t^2}{2}}\,dt$$

SAS Y EL MODELO TOBIT DE REGRESIÓN CENSURADA: PROCEDIMIENTO LIFEREG

El **modelo Tobit de regresión censurada** presenta la forma:

$$y_i = x_i\beta + u_i$$
$$y^*_i = 0 \;\; si\; y_i \le k,\; o\; sea,\; si\;\; x_i\beta + u_i \le k$$
$$y_i^* = y_i \; si\; y_i > k,\; o\; sea,\; si\;\; x_i\beta + u_i > k$$

En este caso se admite que y* sigue una distribución mixta, donde por un lado el valor y*=0 acumula una probabilidad, mientras que en otros valores es continua. La función de probabilidad será:

$$p_i = \begin{cases} p(y_i^* = 0) = p(y_i \le k) = p(x_i\beta + u_i \le k) = p(u_i \le k - x_i\beta) \\ p(y_i^* = y_i) = p(x_i\beta + u_i = y_i) = p(u_i = y_i - x_i\beta) = f(y_i - x_i\beta) \end{cases}$$

Por tanto la función de verosimilitud dependerá de la distribución de la perturbación que normalmente estará en función de la varianza de la perturbación. Maximizando la función de verosimilitud se obtienen estimadores para los parámetros y para la varianza (mediante métodos iterativos).

El **procedimiento LIFERREG** ajusta modelos paramétricos de tiempo de fallos con datos que pueden ser censurados por la izquierda, por la derecha o en un intervalo. En general, el modelo ajustado es de la forma $y = x\beta + \sigma u$ donde la perturbación puede tener distribución normal, logística, exponencial, Weibull, lognormal, gamma o cualquier otra distribución de valores extremos. La variable respuesta y suele ser el logaritmo de tiempos de fallo, σ un parámetro de escala y u la perturbación aleatoria. A veces se utiliza la transformación exponencial formulando el modelo como $T = exp(u)T_o^{\sigma}$. El modelo Tobit se obtiene cuando se considera censura por la izquierda y, como muchas observaciones se sitúan en el valor cero, la expresión del modelo puede ser $y = max(x\beta,\, 0)$.

La variable de interés en el análisis de los modelos de tiempo de fallos (análisis de supervivencia) es la longitud del periodo de tiempo que transcurre desde el principio de algún acontecimiento hasta el final del mismo, o hasta el momento en que ese acontecimiento es observado, lo que puede ocurrir antes de que el acontecimiento acabe. Los datos habitualmente se presentan como un conjunto de duraciones o supervivencias, t1, t2, ..., tn que no necesariamente tienen porqué empezar en el mismo puerto del tiempo.

Una característica inherente al análisis de supervivencia es la **censura**. Se dice que **los datos están censurados** si no se pueden observar por completo. Considérese por ejemplo el análisis del tiempo que transcurre entre el diagnóstico de un determinado tipo de cáncer en un grupo de pacientes y la muerte de los mismos.

Los pacientes son observados cada seis meses, empezando justo en el momento en que se les diagnosticó el cáncer. Por el momento supóngase que a todos los pacientes se les diagnosticó el cáncer el mismo día. Tras seis meses algunos pacientes han muerto y otros no. Para los pacientes sobrevivientes la duración, o supervivencia, es por lo menos igual al periodo observado, ti = 6 meses, pero no es igual a él. Este tipo de censura, la más habitual, se conoce como **censura por la derecha** (tiempo de supervivencia real mayor que el observado). Es posible, así mismo, que exista **censura por la izquierda**, en cuyo caso el tiempo de supervivencia real es menor que el observado. Supongamos por ejemplo que estamos interesados en la supervivencia de un grupo de pacientes con síntomas de un determinado tipo de cáncer, hayan sido o no diagnosticados. En este caso algunos pacientes pueden haber muerto antes de que se les diagnosticase. Tales pacientes presentan censura por la izquierda. La censura también puede ser **censura de intervalo**, por cuanto se conoce que el evento irrepetible ha ocurrido en un intervalo de tiempo determinado. Supongamos ahora que algunos de los pacientes sobrevivientes seis meses después de serles diagnosticado el cáncer, han muerto en la observación, un año después. Existe entonces una censura de intervalo, entre seis meses y año.

La sintaxis del procedimiento LIFEREG es la siguiente:

PROC LIFEREG opciones;
CLASS variables;
MODEL respuesta=independientes / opciones;
BY variables;
OUTPUT OUT=conjunto de datos de salida / estadísticos=nombres;
WEIGHT variable;

Las **opciones de PROC LIFEREG** son: DATA=conjunto de datos de entrada, OUTEST=conjunto de datos que contiene los parámetros estimados y otros estadísticos del ajuste, COVOUT (guarda la matriz de covarianzas estimada), NOPRINT (elimina la salida) y ORDER=DATA | FORMATTED | FREQ | INTERNAL).

Las **opciones de definición de modelo en la sentencia MODEL** son las siguientes:

*label: MODEL respuesta*censor(lista)=independientes / opciones ;*

label: MODEL (inferior, superior)=independientes / opciones;

label: MODEL eventos/tiradas=independientes / opciones;

El primer modelo especifica censura por la derecha en la variable respuesta indicando que cuando la variable censor toma los valores de la lista las observaciones son censuradas.

El segundo modelo indica censura de la variable respuesta en el intervalo (inferior, superior). Si el valor superior no existe o es *missing* se tiene censura por la izquierda y si el valor inferior no existe o es *missing* se tiene censura por la derecha. El tercer modelo especifica dos variables discretas para una respuesta binaria.

Las **opciones de la sentencia MODEL** son: CONVERGE = valor de cambio entre dos estimaciones consecutivas que hace que el algoritmo de convergencia finalice, CORRB (muestra la matriz de correlaciones de los parámetros estimados), COVB (muestra la matriz de covarianzas de los parámetros estimados), DISTRIBUTION=NORMAL | LOGISTIC | EXTREME o EXTREMEVALUE o GOMPERTZ indica la función de distribución utilizada en el modelo de respuesta lineal de probabilidad, NOLOG (no se realiza transformación logarítmica de la variable respuesta), INITIAL=lista de valores iniciales para los parámetros, INTERCEP=valor (inicializa el parámetro constante), SCALE=valor inicial del parámetro de escala, NOSCALE (toma parámetro de escala fijo), ITPRINT (muestra la historia de iteraciones), SHAPE1= valor inicial del primer parámetro de forma, NOSHAPE1 (toma parámetro de forma fijo), MAXITER=máximo número de iteraciones, NOINT (ajusta el modelo sin constante) y SINGULAR=valor del criterio de singularidad.

Los **estadísticos a utilizar OUTPUT** son: CENSORED (variable que toma el valor 1 para observaciones censuradas y el valor 0 para observaciones no censuradas) CDF (variable que contiene la función de distribución evaluada en la respuesta observada), CONTROL (variable del conjunto inicial de datos para el control de la estimación de cuantiles que vale 1 para cada cuantil a estimar), PREDICTED | P (variable que contiene los cuantiles estimados y si la variable respuesta es binomial la variable vale $1-F(-X\beta)$), QUANTILES | QUANTILE | Q (lista de valores para los que se calculan los cuantiles), STD (variable que contiene el error estándar de las estimaciones $X\beta$) y XBETA (estimaciones de $X\beta$). En un **modelo Tobit** $y_i = x_i\beta + u_i$, XBETA estima $x_i\beta$ y los valores predichos vienen dados por:

$$E(Y_i) = \Phi\left(\frac{X_i'\beta}{\sigma}\right)(X_i'\beta + \sigma\lambda_i) \qquad \lambda_i = \frac{\phi(X_i'\beta/\sigma)}{\Phi(X_i'\beta/\sigma)}$$

Φ y ϕ son respectivamente la función de densidad y la función de distribución.

Como primer ejemplo supongamos que disponemos de las variables: *horas* que indica el número de horas que las mujeres casadas trabajan fuera de su hogar en un año dado, *educacion* que indica sus años de educación y *experiencia* que indica los años de experiencia en el trabajo. Vamos a considerar un modelo Tobit que ajuste las horas trabajadas con educación y experiencia como covariables. El valor cero de la variable *horas* indica que la mujer está en paro, con lo que habrá censura por la izquierda, que se indicará considerando intervalos que contengan las observaciones censuradas de la variable horas de modo que su extremo inferior sea desaparecido y su extremo superior sea cero (valor de censura).

```
data conjunto;
      input Horas Educacion Experiencia @@;
      if horas eq 0
         then inferior=.;
         else inferior=Horas;
   datalines;
   0 8 9 0 8 12 0 9 10 0 10 15 0 11 4 0 11 6
   1000 12 1 1960 12 29 0 13 3 2100 13 36
   3686 14 11 1920 14 38 0 15 14 1728 16 3
   1568 16 19 1316 17 7 0 17 15
  ;
proc lifereg data=conjunto outest=OUTEST(keep=_scale_);
      model (inferior, horas) = educacion experiencia / d=normal;
      output out=OUT xbeta=Xbeta;
   run;
```

La salida del ajuste es la siguiente:

```
                    The LIFEREG Procedure

                     Model Information

                Data Set                     WORK.CONJUNTO
                Dependent Variable               inferior
                Dependent Variable                  Horas
                Number of Observations                 17
                Noncensored Values                      8
                Right Censored Values                   0
                Left Censored Values                    9
                Interval Censored Values                0
                Name of Distribution               NORMAL
                Log Likelihood                 -74.9369977

          Algorithm converged.

                  Analysis of Parameter Estimates

                              Standard
  Variable      DF   Estimate      Error Chi-Square Pr > ChiSq Label

  Intercept      1    -5598.6     2850.2     3.8583     0.0495 Intercept
  Educacion      1  373.14771  191.88717     3.7815     0.0518
  Experiencia    1   63.33711   38.36317     2.7258     0.0987
  Scale          1     1582.9  442.67318                        Normal scale
```

Se obtienen estimaciones para los parámetros que son significativas al 90%.

MODELOS DE VARIABLE DEPENDIENTE LIMITADA CON EVIEWS: MLP, LOGIT Y PROBIT

Inicialmente ajustaremos un modelo lineal de probabilidad a los datos de 753 mujeres casadas en 1975 (archivo *casadas.wf1*), de las cuales 428 participaron en el mercado laboral en algún momento del año (*inlf*=1) y 325 no participaron (*inlf*=0).

Se considera que la probabilidad de participar en el mercado de trabajo (*inlf*) depende de otras fuentes de ingresos (*nwifeinc*), de los años de educación (*educ*), de los años de experiencia laboral (*exper* y *exper*^{2}), de la edad (*age*), del número de hijos de menos de 6 años (*kidslt6*) y del número de hijos entre 6 y 18 años (*kidsg6*).

Mediante un modelo lineal de probabilidad analizaremos la dependencia de la probabilidad de participación de la mujer en el mercado de trabajo en función de las variables explicativas anteriormente definidas.

Plantearemos el siguiente modelo lineal de probabilidad (MLP):

$$Inlf = \beta_0 + \beta_1\, nwifeinc + \beta_2\, educ + \beta_3\, exper + \beta_4\, exper^2 + \beta_5\, age + \beta_6\, kidslt6 + \beta_7\, kidsg6$$

Comenzaremos realizando la ***estimación del MLP*** mediante MCO. Para ello se elige *Quick* → *Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *LS-Least Squares (NLS and ARMA)* en el campo *Method* para ajustar por mínimos cuadrados ordinarios (Figura 5-38) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 5-39 con buenas significatividades individuales y conjuntas para los parámetros estimados (salvo quizá *kidsge6*). El R^2 y el estadístico de Durbin Watson no son muy buenos, pero tenemos que pensar que estamos estimando un modelo lineal de probabilidad.

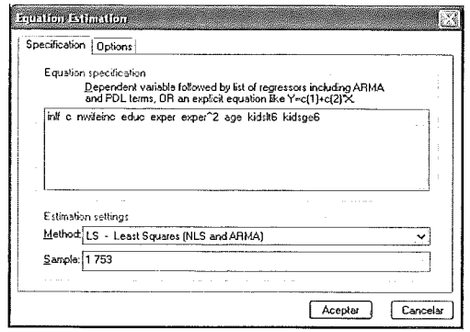


Figura 5-38

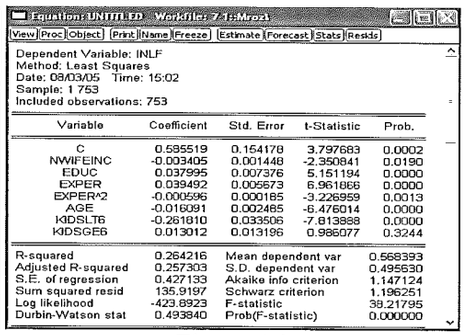


Figura 5-39

Pero el problema relevante en un MLP es la heteroscedasticidad. Mediante *View* → *Residual Tests* → *White Heteroskedasticity* (Figura 5-40) se obtienen p-valores de F y Chi menores que 0,05 (Figura 5-41), luego existe heteroscedasticidad. Utilizaremos entonces para la estimación un método robusto a la heteroscedasticidad, por ejemplo el de White. La estimación mediante el método de White se hará ahora rellenando la solapa *Options* de la pantalla *Equation Estimation* como se indica en la Figura 5-42. Al hacer clic en *Aceptar* se obtiene la estimación sin heteroscedasticidad de la Figura 5-43.

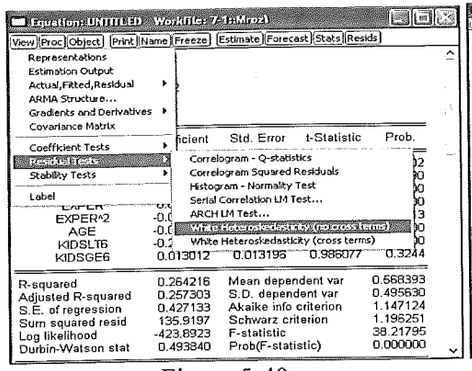


Figura 5-40

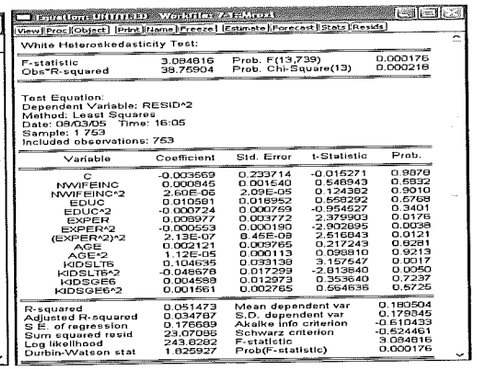


Figura 5-41

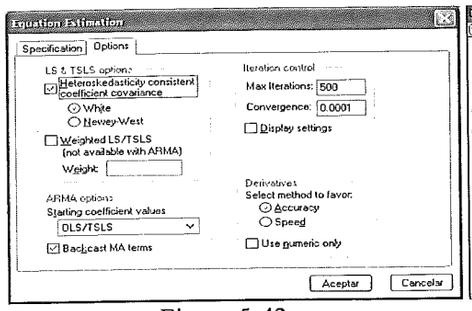


Figura 5-42

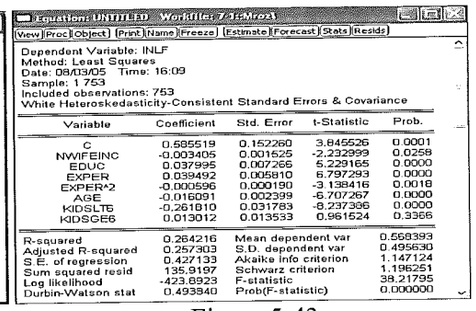


Figura 5-43

Salvo *kidsge6* todas las variables son estadísticamente significativas y tienen los signos esperados. También se observa una fuerte significatividad conjunta. Al interpretar los coeficientes se observa que si aumenta la educación en un año (y las restantes variables permanecen constantes), la probabilidad de la mujer de participar en el mercado de trabajo aumenta en 0,038. El efecto de otras fuentes de ingresos (*nwifeinc*) es significativo, pero muy pequeño, ya que si otros ingresos aumentan en una unidad, la probabilidad de la mujer de participar en el mercado de trabajo disminuye en 0,0034 (con el resto de los factores constantes). Al variar la experiencia en un año, con el resto de los factores fijos, la variación estimada en la probabilidad de la mujer de participar en el mercado de trabajo viene dada por 0,039-2(0,0006)*exper*. El punto en el que la experiencia no afecta a la probabilidad de participar (y a partir del cual cambia de signo la pendiente) es 0,039/0,0012=32,5 (sólo hay 13 mujeres con *exper*>32). Al aumentar la edad en un año (con todo lo demás constante), la probabilidad de la mujer de participar en el mercado laboral se reduce en 0,016. Tener un hijo más de menos de 6 años, reduce la probabilidad de la mujer de participar en el mercado de trabajo en 0,262, para unos niveles dados de las demás variables. Tener un hijo entre 6 y 18 años, aumenta la probabilidad de la mujer de participar en el mercado de trabajo en 0,013, para niveles fijos del resto de las variables.

Uno de los mayores problemas del MLP es que para ciertas combinaciones de valores de las variables explicativas, las probabilidades estimadas pueden ser menores que 0 o mayores que 1. Podemos calcular para todas las mujeres de la muestra:

$$\hat{P} = 0{,}586 - 0{,}0034 nwifeinc + 0{,}038 educ + 0{,}039 \exp er$$
$$- 0{,}0006 \exp er^2 - 0{,}016 age - 0{,}262 kidslt6 + 0{,}013 kidsge6$$

Para ello hacemos clic en *Forecast* y rellenamos la pantalla de predicción como se indica en la Figura 5-44 (la variable *inlff* contiene los valores de $\hat{P}$ para las 753 mujeres de la muestra). Al pulsar *Aceptar* se obtiene la Figura 5-45 que representa $\hat{P}$, observándose que hay valores superiores a la unidad e inferiores a 0). De hecho, si hacemos doble clic sobre la variable *inlff* en la pantalla Workfile, vemos sus valores. En la Figura 5-46 se observan algunos valores superiores a la unidad, en la Figura 5-47 se observan valores menores que cero. En total hay 17 valores de $\hat{P}$ mayores que uno y 16 menores que cero.

Otra incongruencia del MLP es que el efecto sobre la probabilidad sea el mismo para todos los valores de las variables explicativas.

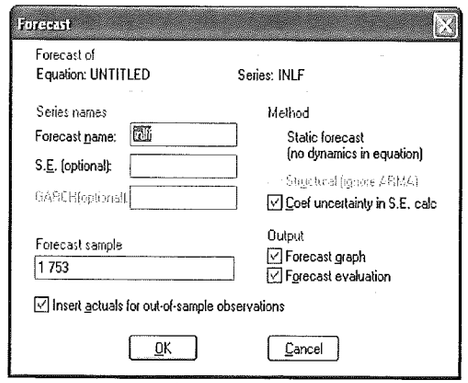


Figura 5-44

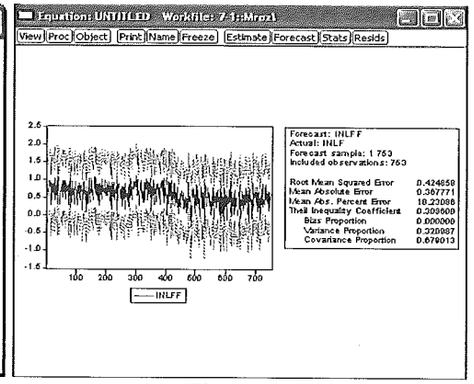


Figura 5-45

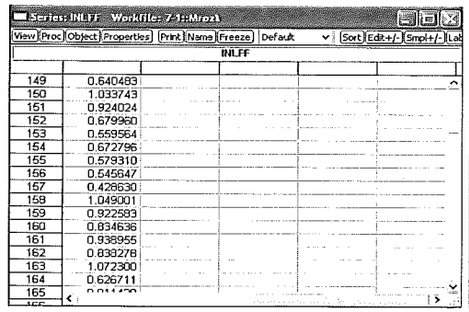


Figura 5-46

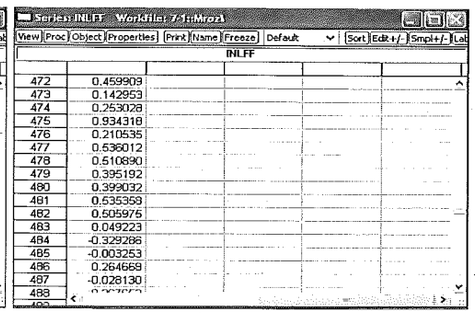


Figura 5-47

A continuación, dadas las posibles deficiencias del modelo lineal de probabilidad, resolvemos el problema anterior utilizando un modelo Probit y un modelo Logit. Adicionalmente se cuantificarán e interpretarán los efectos de las variables explicativas sobre la probabilidad de la mujer de participar en el mercado de trabajo de la forma adecuada

Plantearemos en primer lugar el modelo Probit:

$$Inlf = \Phi(\beta_0 + \beta_1\, nwifeinc + \beta_2\, educ + \beta_3\, exper + \beta_4\, exper^2 + \beta_5\, age + \beta_6\, kidslt6 + \beta_7\, kidsg6)$$

donde Φ es la función de distribución de una normal (0,1).

Comenzaremos realizando la ***estimación mediante el modelo Probit***. Para ello se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *BINARY-Binary choice (logit, probit, extreme value)* en el campo *Method,* se elige *Probit* en el campo *Binary estimation method* para realizar el ajuste mediante el modelo Probit (Figura 5-48) y se hace clic en *Aceptar*.

Se obtienen los resultados de la Figura 5-49 con buenas significatividades individuales para los parámetros estimados (salvo la constante y *kidsge6*). La significatividad conjunta es muy alta porque el p-valor del estadístico de la razón de verosimilitud es muy pequeño. El Pseudo R^2 de McFadden no se acerca demasiado a la unidad (0,22). Los valores de los criterios de información (Akaike, Schwarz y Hannan-Quinn) son adecuados porque son bajos y muy parecidos.

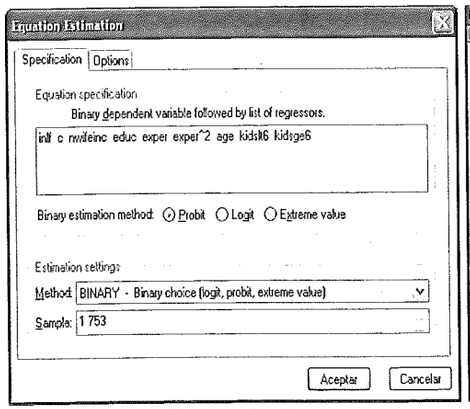


Figura 5-48

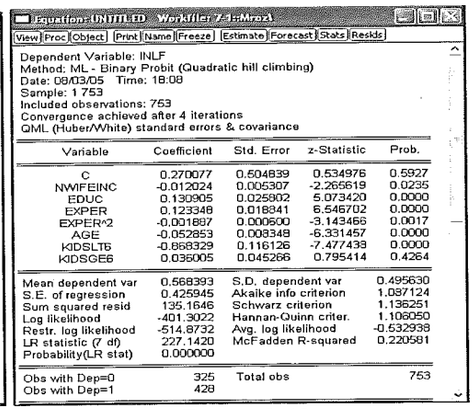


Figura 5-49

Otro criterio para medir la bondad del ajuste del modelo Probit es el criterio del porcentaje de predicciones correctas que consiste en observar el porcentaje de veces en que el valor de Y_i observado coincide con su predicción. Esta tarea se lleva a cabo con Eviews mediante *View → Expectation-Prediction Table* (Figura 5-50).

Al hacer clic en *Aceptar* se obtiene la salida de la Figura 5-51 en la que se observa que el modelo predice adecuadamente el 73,44% de las observaciones. Se predicen mejor los unos (la participación de la mujer en el mercado de trabajo) con un 81,31% de aciertos frente a un 63,08% de la no participación.

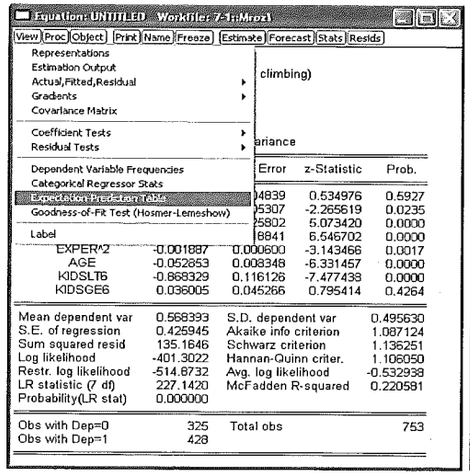


Figura 5-50

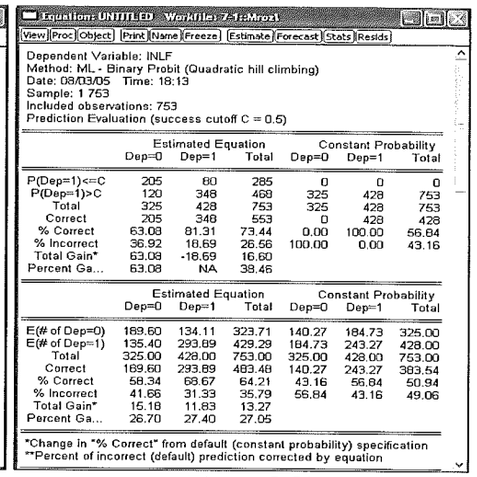


Figura 5-51

Para cuantificar e interpretar los efectos de las variables explicativas sobre la probabilidad de la mujer de participar en el mercado de trabajo tendremos presente que el efecto parcial de una variable explicativa continua X_j sobre la probabilidad de respuesta $P(Y=1|X)$ es:

$$\frac{\partial P(Y=1|\mathbf{X})}{\partial X_j} = g(\mathbf{X}\beta)\beta_j$$

Como para cada observación tenemos un efecto, evitaremos el cálculo de 753 efectos calculando los efectos para las observaciones medias. Para ello calculamos las medias de las variables del modelo seleccionándolas en la pantalla *Workfile* y abriéndolas como un grupo con el botón secundario del ratón mediante *Open → as Group* (Figura 5-52). A continuación se elige *View → Descriptive Statistics→ Individual samples* (Figura 5-53) y se obtienen los estadísticos descriptivos para todas las variables del modelo, incluida la media (Figura 5-54).

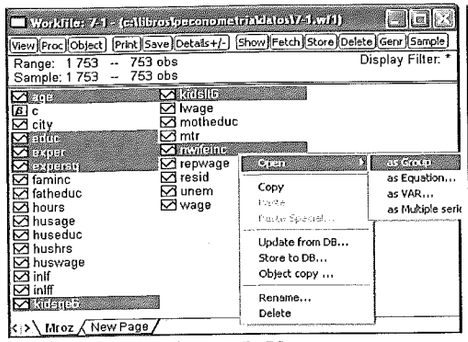


Figura 5-52

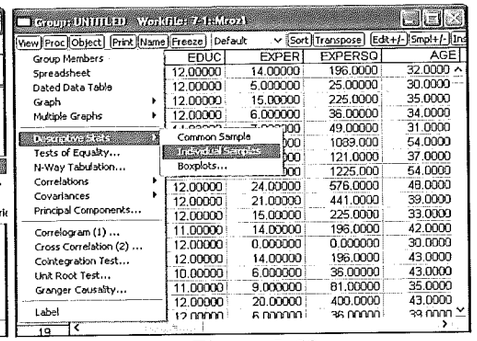


Figura 5-53

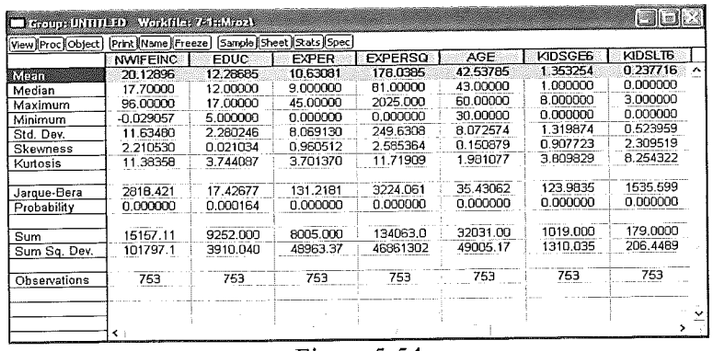


Figura 5-54

Tenemos:

$$\overline{nwifeinc} = 20{,}13 \quad \overline{educ} = 12{,}29 \quad \overline{\exp er} = 10{,}63$$

$$\overline{age} = 42{,}54 \quad \overline{kidslt6} = 0{,}24 \quad \overline{kidsge6} = 1{,}35$$

con lo que:

$$g(\overline{\mathbf{X}}\hat{\beta}) = \phi(\hat{\beta}_0 + \hat{\beta}_1\overline{nwifeinc} + \hat{\beta}_2\overline{educ} + \hat{\beta}_3\overline{\exp er} + \hat{\beta}_4\overline{\exp er^2} +$$
$$+ \hat{\beta}_5\overline{age} + \hat{\beta}_6\overline{kidslt6} + \hat{\beta}_7\overline{kidsge6}) = 0{,}391$$

ϕ = Función de densidad de la normal (0,1)

Podemos obtener el efecto parcial estimado de las variables continuas, para los valores medios de X, multiplicando los coeficientes estimados del modelo Probit por 0,391.

Por tanto, un año adicional de educación aumenta la probabilidad de participar en el mercado de trabajo en 0,131x0,391 = 0,05. Al aumentar la edad en un año la probabilidad de participar se reduce en 0,02 (-0,053x0,391=-0,021). Al aumentar la experiencia en un año, la probabilidad de participar varía en [0,123-(2x0,0019x*exper*]x0,391. Un aumento unitario en los otros ingresos hace que la probabilidad de participar se reduzca en 0,005 (-0,012x0,391 = -0,05).

También podemos calcular el efecto estimado al variar una variable discreta. Por ejemplo, para *kidslt6* el efecto sobre la probabilidad de trabajar de pasar de tener 0 a 1 hijo (menor de 6 años) es una reducción en 0,336 porque siendo Φ = Función de distribución de la normal (0,1) se tiene:

$$\Phi(\hat{\beta}_0 + \hat{\beta}_1\overline{nwifeinc} + \hat{\beta}_2\overline{educ} + \hat{\beta}_3\overline{\exp er} + \hat{\beta}_4\overline{\exp er^2} + \hat{\beta}_5\overline{age} + \hat{\beta}_7\overline{kidsge6}) -$$
$$\Phi(\hat{\beta}_0 + \hat{\beta}_1\overline{nwifeinc} + \hat{\beta}_2\overline{educ} + \hat{\beta}_3\overline{\exp er} + \hat{\beta}_4\overline{\exp er^2} + \hat{\beta}_5\overline{age} + \hat{\beta}_6 + \hat{\beta}_7\overline{kidsge6}) =$$
$$= 0{,}660 - 0{,}324 = 0{,}336$$

y el efecto sobre la probabilidad de trabajar de la mujer de pasar de 1 a 2 hijos (menor de 6 años) se reducirá en 0,231porque:

$$\Phi(\hat{\beta}_0 + \hat{\beta}_1\overline{nwifeinc} + \hat{\beta}_2\overline{educ} + \hat{\beta}_3\overline{\exp er} + \hat{\beta}_4\overline{\exp er^2} + \hat{\beta}_5\overline{age} + \hat{\beta}_6 + \hat{\beta}_7\overline{kidsge6}) -$$
$$\Phi(\hat{\beta}_0 + \hat{\beta}_1\overline{nwifeinc} + \hat{\beta}_2\overline{educ} + \hat{\beta}_3\overline{\exp er} + \hat{\beta}_4\overline{\exp er^2} + \hat{\beta}_5\overline{age} + 2\hat{\beta}_6 + \hat{\beta}_7\overline{kidsge6}) =$$
$$= 0{,}324 - 0{,}093 = 0{,}231$$

Ahora realizaremos la ***estimación mediante el modelo Logit***. Para ello se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *BINARY-Binary choice (logit, probit, extreme value)* en el campo *Method,* se elige *Logit* en el campo *Binary estimation method* para realizar el ajuste mediante el modelo Probit (Figura 5-55) y se hace clic en *Aceptar*.

Se obtienen los resultados de la Figura 5-56 con buenas significatividades individuales para los parámetros estimados (salvo la constante y *kidsge6*) al igual que en el caso del modelo Probit. La significatividad conjunta es muy alta porque el p-valor del estadístico de la razón de verosimilitud es muy pequeño. El Pseudo R^2 de McFadden no se acerca demasiado a la unidad (0,219). Los valores de los criterios de información (Akaike, Schwarz y Hannan-Quinn) son adecuados.

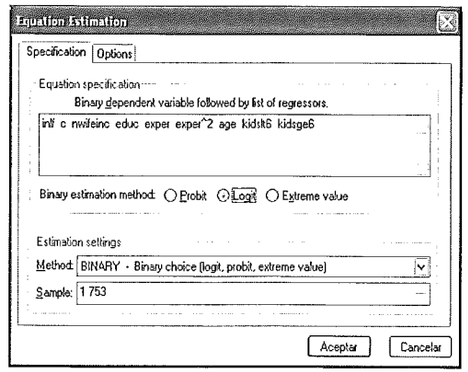


Figura 5-55

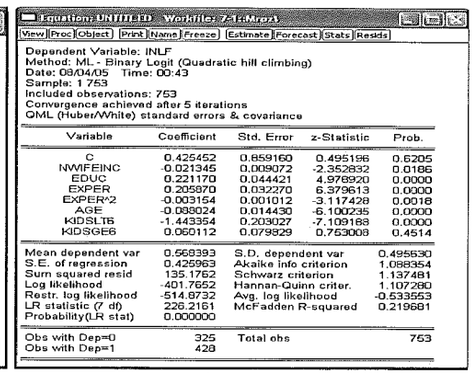


Figura 5-56

Para la función logística:

$$g(\overline{X}\hat{\beta}) = \lambda(\overline{X}\hat{\beta}) = \frac{e^{\overline{X}\hat{\beta}}}{(1-e^{\overline{X}\hat{\beta}})^2} = \Lambda(\overline{X}\hat{\beta})(1-\Lambda(\overline{X}\hat{\beta})) = 0{,}227$$

(λ es la función de densidad de la logística y Λ su función de distribución)

y al calcular los efectos parciales medios de las variables multiplicando los coeficientes estimados por 0,227, se obtienen los mismos resultados que para el modelo Probit.

MODELOS DE RECUENTO CON EVIEWS: POISSON, BINOMIAL NEGATIVA Y EXPONENCIAL

Como ejemplo consideramos el archivo *arrestos.wf*1 que contiene información sobre la variable *narr*86 que representa el número de veces que es arrestado un hombre durante 1986 y que se desea explicar mediante las variables *pcnv* (proporción de arrestos previos en los que fue condenado), *avgsen* (duración media en meses de las sentencias), *tottime* (tiempo en prisión en meses), *ptime*86 (tiempo en prisión en 1986 en meses), *qemp*86 (trimestres empleados en 1986), *inc*86 (ingresos legales en 1986 en cientos de dólares), *black* (ficticia que vale 1 para arrestado de raza negra), *hispan* (ficticia que vale 1 para arrestado de raza hispana) y *born*60 (ficticia que vale 1 si el arrestado nació en 1960). Como la variable explicada toma el valor 0 en un número elevado de casos y sólo toma valor superior a 5 en 8 casos, podría utilizarse un modelo de datos de recuento. Estudiaremos el efecto de las variables explicativas sobre la explicada utilizando modelos de datos de recuento de Poissson, exponencial y binomial negativa.

Realizamos en primer lugar la ***estimación del modelo mediante un modelo de recuento de Poisson***. Para ello se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *COUNT-Integer count data* en el campo *Method,* se señala *Poisson* (*ML and QML*) en el campo *Count estimation method* para ajustar por un modelo de datos de recuento de Poisson (Figura 5-57) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 5-58 con buenas significatividades individuales y conjunta para los parámetros estimados (salvo para *avgsen, tottime, ptime86* y *born*60). El Pseudo R^2 es bastante pequeño y los valores de los criterios de información son bastante aceptables. Se observan parámetros de ajuste similares al caso anterior. *Mediante View → Representations* (Figura 5-59) se obtienen las ecuaciones del modelo ajustado (Figura 5-60).

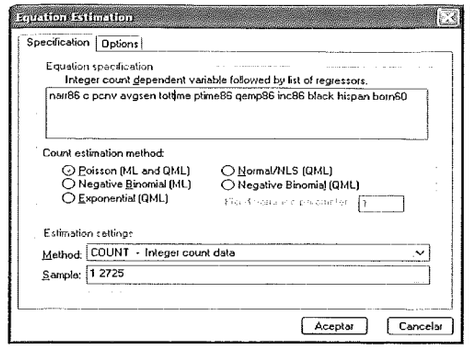


Figura 5-57

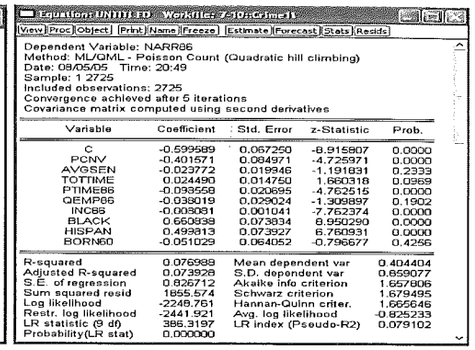


Figura 5-58

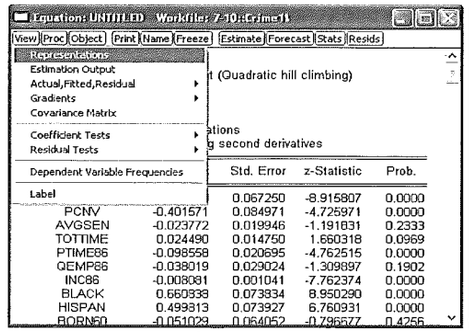


Figura 5-59

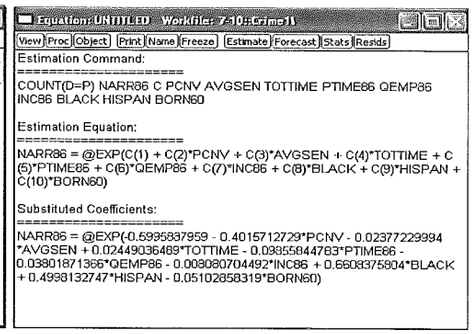


Figura 5-60

Como el ajuste ahora es $Ln(narr86) = X\beta + u \Leftrightarrow narr86 = e^{X\beta + u}$, a la hora de interpretar los resultados vemos que si el individuo está empleado en 1986 un trimestre más (el resto de factores fijo), el número de arrestos esperado en ese año se reduce en un 3,8% (no significativo). Con todo lo demás igual, el número de arrestos esperado de un hombre negro es un 66% mayor que el de un blanco y el de un hispano es un 50% superior al de un blanco. Los demás parámetros se interpretarían de forma similar.

Si en la pantalla *Equation Estimation* hacemos clic en la solapa *Options* y elegimos *Robust Covariances-GLM* (Figura 5-61), al hacer clic en *Aceptar* obtenemos la estimación del modelo de recuento de Poisson con errores estándar GLM calculados bajo $V(Y|X) = \sigma^2 E(Y|X)$ según se muestra en la Figura 5-62. Se observa que los resultados no difieren demasiado del caso anterior y las estimaciones de los parámetros son muy parecidas. Obsérvese que la estimación de la varianza es:

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\frac{\hat{u}_i^2}{\hat{Y}_i} = 1{,}516788154$$

Como $\sigma^2 > 1$ tenemos un caso de *sobredispersión* respecto a la varianza de Poisson. *Mediante View → Representations* (Figura 5-63) se obtiene el modelo ajustado (Figura 5-64).

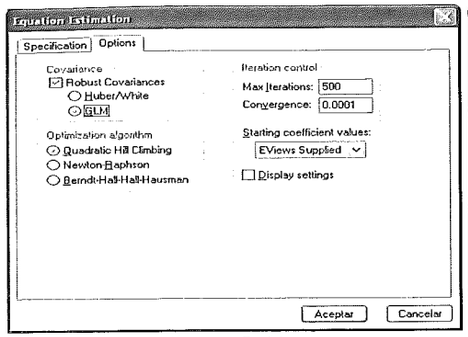


Figura 5-61

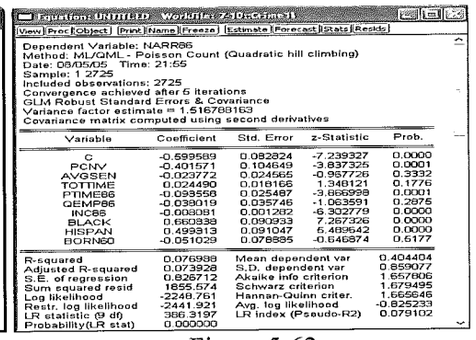


Figura 5-62

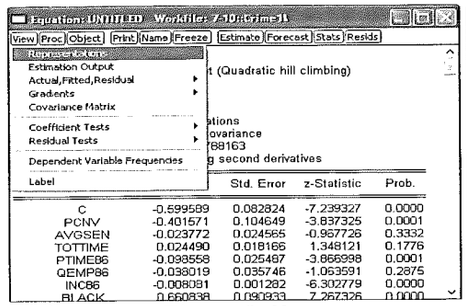


Figura 5-63

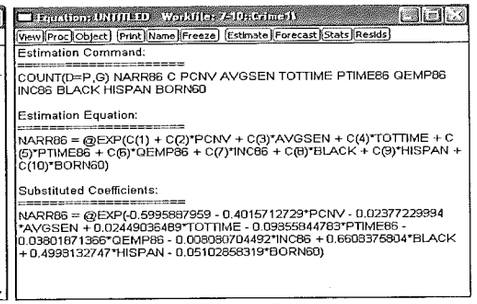


Figura 5-64

Realizamos ahora la ***estimación del modelo mediante un modelo de recuento de binomial negativa (con cuasi máxima verosimiltud)***. Para ello se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *COUNT-Integer count data* en el campo *Method,* se señala *Negative Binomial* (*QML*) en el campo *Count estimation method* para ajustar por un modelo de datos de recuento de binomial negativa con máxima verosimilitud cuadrática (Figura 5-65) y se hace clic en *Aceptar*.

Se obtienen los resultados de la Figura 5-66 con buenas significatividades individuales y conjunta para los parámetros estimados (salvo para *avgsen, tottime* y *born*60). El Pseudo R^2 es bastante pequeño y los valores de los criterios de información son bastante aceptables. Además se observan resultados muy similares a los de los métodos anteriores. Mediante *View → Representations* (Figura 5-67) se obtienen las ecuaciones del modelo ajustado (Figura 5-68).

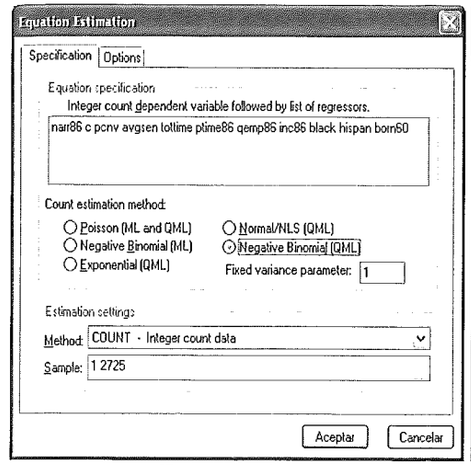


Figura 5-65

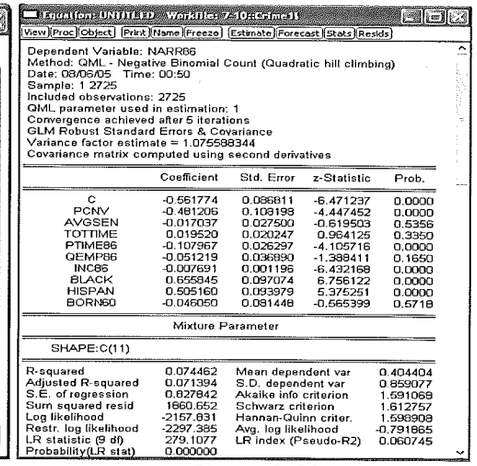


Figura 5-66

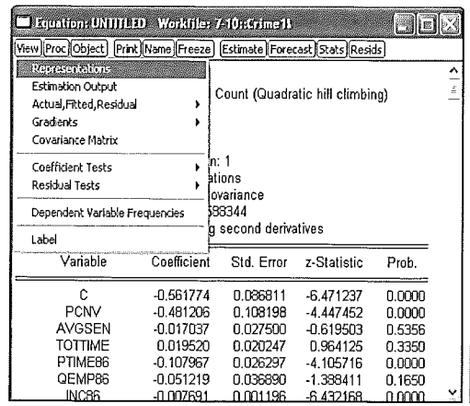


Figura 5-67

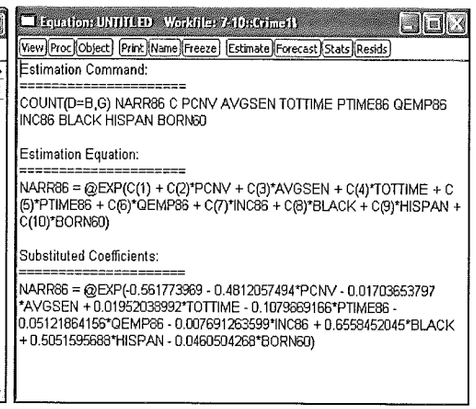


Figura 5-68

Por último, realizamos ahora la ***estimación del modelo mediante un modelo de recuento exponencial (con cuasi máxima verosimiltud)***. Para ello se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *COUNT-Integer count data* en el campo *Method,* se señala *Exponential* (*QML*) en el campo *Count estimation method* para ajustar por un modelo de datos de recuento exponencial con máxima verosimilitud cuadrática (Figura 5-69) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 5-70. Mediante *View → Representations* (Figura 5-71) se obtienen las ecuaciones del modelo ajustado (Figura 5-72).

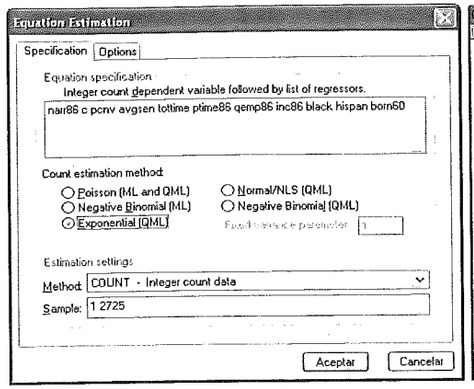


Figura 5-69

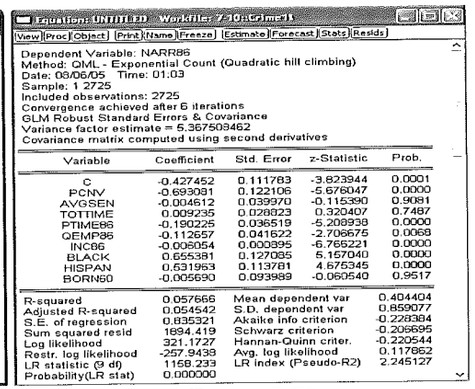


Figura 5-70

Se observan buenas significatividades individuales y conjunta para los parámetros estimados (salvo para *avgsen, tottime* y *born*60). El Pseudo R^2 es grande y los valores de los criterios de información son bastante aceptables. Además se observan resultados muy similares a los de los métodos anteriores.

El modelo de datos de recuento más adecuado para este ajuste es el modelo exponencial, ya que es el que menores valores presenta de los criterios de información de Akaike, Schwarz y Hannan-Quinn.

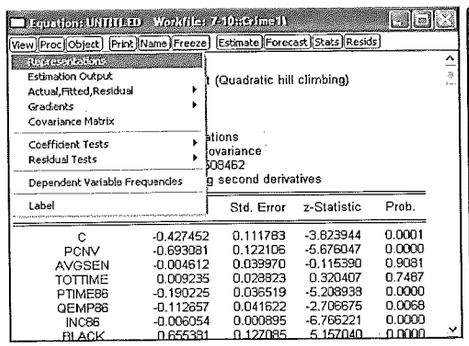


Figura 5-71

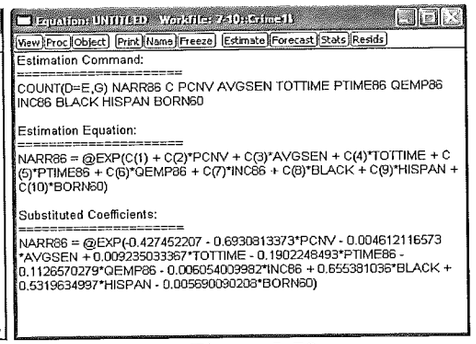


Figura 5-72

MODELOS TOBIT CENSURADO Y TRUNCADO CON EVIEWS. MÉTODO DE HECKMAN Y RATIO DE MILLS

Consideramos datos procedentes de un informe de una empresa sobre 753 familias (archivo *familias.wf*1) en el que se estimó un modelo que relacionaba el gasto en adquisición de un automóvil (*GASTO*) con la renta familiar en el último año (*RENTA*), el número de hijos inferiores a 18 años (*HIJOS*) y la edad del cabeza de familia (*EDAD*). Se especificó a la siguiente ecuación de comportamiento:

$$GASTO_i = \beta_0 + \beta_1\, RENTA_i + \beta_2\, HIJOS_i + \beta_3\, EDAD_i + u_i$$

Se trata de estimar el modelo anterior teniendo presente que puede haber individuos encuestados que no hayan comprado vehículo en ese período, en cuyo caso se les ha asignado un gasto nulo. La tarea es hallar los efectos marginales de las variables explicativas sobre la explicada y compararlos con las estimaciones MCO de los coeficientes del modelo y calcular la elasticidad demanda renta en el punto medio.

Estamos ante el caso particular de censura de datos que se presenta cuando la variable dependiente vale cero para una parte de la población y es continua para los valores positivos. Estamos entonces ante modelo de variable limitada con solución de esquina. Concretamente ante un modelo Tobit censurado con solución de esquina.

Comenzaremos realizando la ***estimación mediante el modelo Tobit censurado por máxima verosimilitud***. Para ello se elige *Quick* → *Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *CENSORED-Censored or truncated data (Tobit)* en el campo *Method,* se sitúa en cero en el campo *Left* y se señala *Actual censoring value* porque el modelo Tobit con solución de esquina está censurado a la izquierda por el valor cero, se elige *Normal* porque consideramos el modelo con distribución normal (Figura 5-73) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 5-74 con buenas significatividades individuales para los parámetros estimados. El R^2 está muy lejano de la unidad, pero esto es habitual en este tipo de modelos. Se observa que de un total de 753 observaciones hay 336 valores nulos (observaciones no censuradas, o sea, individuos encuestados que no habían comprado vehículo en ese año) y 417 observaciones censuradas a la izquierda (porque son valores positivos con el valor de censura situado a su izquierda). Los valores de los criterios de información no son demasiado elevados y son bastante coincidentes entre ellos.

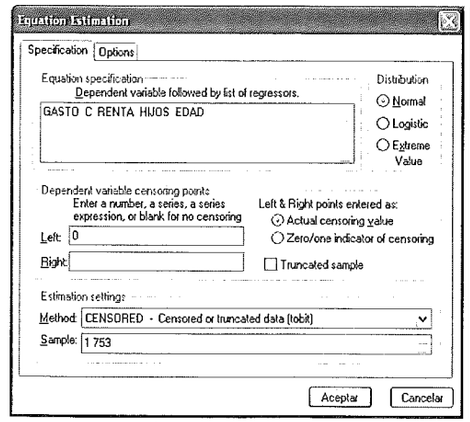


Figura 5-73

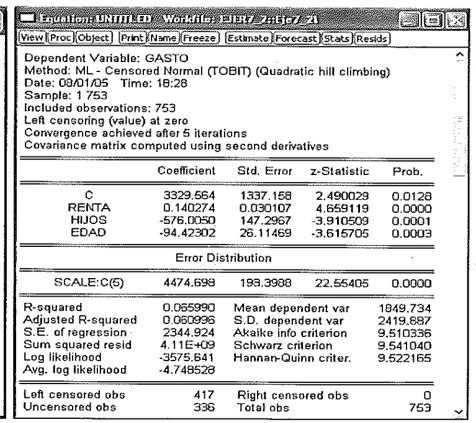


Figura 5-74

El efecto marginal en el modelo Tobit censurado de cada una de las variables se obtiene mediante la expresión (aplicada en los valores medios):

$$\frac{\delta E(Y|\mathbf{X})}{\delta X_j} = \beta_j \Phi(\mathbf{X}\beta / \sigma)$$

A partir del vector de las medias de todas las variables explicativas $\overline{X}_i$ y del vector de los parámetros estimados $\hat{\beta}$, se tiene que:

$$\frac{\overline{X}_i\hat{\beta}}{\hat{\sigma}} = \frac{-228{,}486}{447{,}698} = -0{,}051 \Rightarrow \Phi\left(\frac{\overline{X}_i\hat{\beta}}{\hat{\sigma}}\right) = \Phi(-0{,}051) = \int_0^{-0{,}051} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} ds = 0{,}4796$$

Los ***efectos marginales para el punto medio*** pueden calcularse entonces como sigue:

$$\frac{\partial E(GASTO_i)}{\partial RENTA_i} = \hat{\beta}_1 \Phi\left(\frac{\overline{X}_i\hat{\beta}}{\hat{\sigma}}\right) = 0{,}140274 * 0{,}4796 = 0{,}06728$$

$$\frac{\partial E(GASTO_i)}{\partial HIJOS_i} = \hat{\beta}_2 \Phi\left(\frac{\overline{X}_i\hat{\beta}}{\hat{\sigma}}\right) = -576{,}005 * 0{,}4796 = -276{,}288$$

$$\frac{\partial E(GASTO_i)}{\partial EDAD_i} = \hat{\beta}_3 \Phi\left(\frac{\overline{X}_i\hat{\beta}}{\hat{\sigma}}\right) = -94{,}42302 * 0{,}4796 = -45{,}291$$

Por tanto, la ***elasticidad demanda renta*** se estimará a partir de nuestro modelo Tobit como sigue:

$$\frac{\dfrac{\partial E(GASTO_i)}{\overline{GASTO}}}{\dfrac{\partial RENTA_i}{\overline{RENTA}}} = \frac{\partial E(GASTO_i)}{\partial RENTA_i}\frac{\overline{RENTA}}{\overline{GASTO}} = \hat{\beta}_1 \Phi\left(\frac{\overline{X}_i\hat{\beta}}{\hat{\sigma}}\right)\frac{\overline{RENTA}}{\overline{GASTO}} = 0{,}06728\frac{587{,}73}{21{,}031} = 0{,}42$$

Para poder comparar los efectos marginales para el punto medio con los coeficientes estimados del modelo lineal mediante MCO realizaremos esta última estimación. Para ello se elige *Quick* → *Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *LS-Two stage Least Squares (NLS and ARMA)* en el campo *Method* para ajustar por mínimos cuadrados (Figura 5-75) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 5-76.

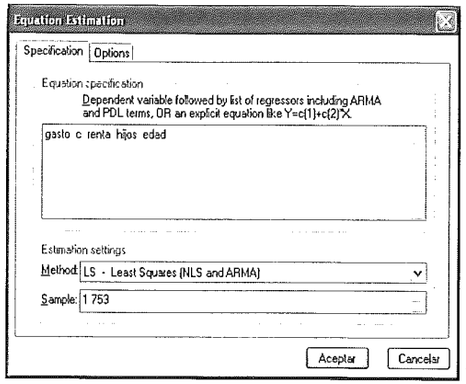


Figura 5-75

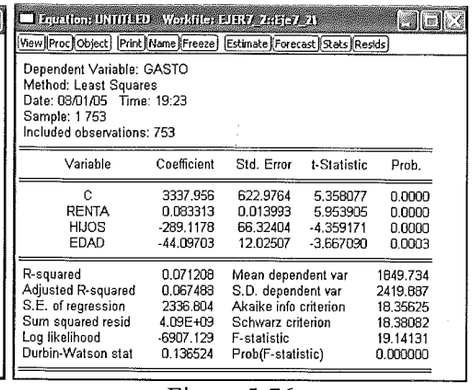


Figura 5-76

Se observa que los efectos marginales del modelo Tobit se acercan pero no coinciden con los valores estimados del modelo MCO, quedando siempre un poquito por debajo. La diferencia entre estas estimaciones MCO y los efectos marginales del Tobit serían los errores que se cometerían en las estimaciones de los parámetros del modelo si se utilizara MCO en vez de un modelo Tobit.

Se trata ahora de estimar una ecuación salarial especificando un modelo econométrico en el que se relaciona el logaritmo de los salarios percibidos por los empleados Y con la experiencia laboral (X_2), su cuadrado (X_3), el nivel educativo (X_4), el estado civil del trabajador (X_5) y la titulación universitaria o no del mismo (X_6) a través de la siguiente ecuación de comportamiento:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + u_i$$

Realizaremos la estimación de esta ecuación de comportamiento salarial utilizando un modelo Tobit truncado y calcularemos los efectos marginales de las variables X_2, X_3 y X_4 sobre el salario. Los datos se recogen en el archivo *salarios.wf1*.

Dado que sólo observamos la oferta salarial para los individuos que están trabajando (y no para los que no trabajan) estamos ante un caso de selección muestral no aleatoria en el que observamos Y o no dependiendo de otra variable (el empleo). Estamos claramente ante un caso de ***truncamiento*** en el que se observa la oferta salarial dependiendo de otra variable que es la oferta de trabajo.

Como estamos en un ***modelo Tobit truncado*** utilizaremos el menor valor de la variable dependiente Y como valor de truncamiento. Este valor mínimo es el que queda más a la izquierda de los restantes valores de Y. Para calcular el valor mínimo de Y hacemos doble clic sobre Y en la pantalla *Workfile* y elegimos *View* → *Descriptive Statistics* → *Stats Table* (Figura 5-77). En la Figura 5-78 observamos que el valor mínimo de Y es 1,8718.

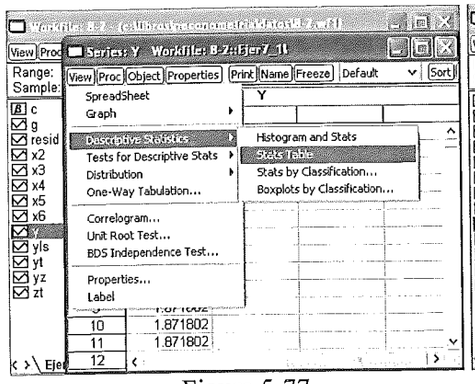


Figura 5-77

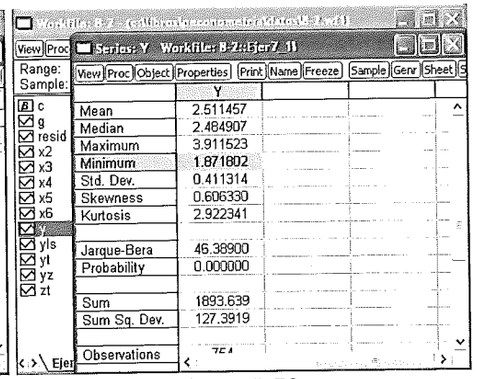


Figura 5-78

Realizamos ahora la ***estimación de la ecuación de comportamiento salarial mediante el modelo Tobit truncado por máxima verosimilitud***. Para ello se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *CENSORED-Censored or truncated data (Tobit)* en el campo *Method,* se sitúa 1,8718 en el campo *Left* y se señala *Actual censoring value* porque el modelo Tobit truncado está censurado a la izquierda por dicho valor, se elige *Normal* porque consideramos el modelo con distribución normal, se señala *Truncated sample* (Figura 5-79) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 5-80 con buenas significatividades individuales para los parámetros estimados (salvo quizás X_3 y X_5 al 80% y X_6 al 50%). El R^2 está muy lejano de la unidad, pero esto es habitual en este tipo de modelos. Los valores de los criterios de información no son demasiado elevados y son bastante coincidentes entre ellos.

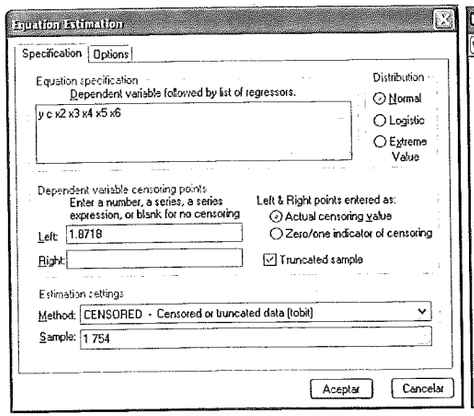


Figura 5-79

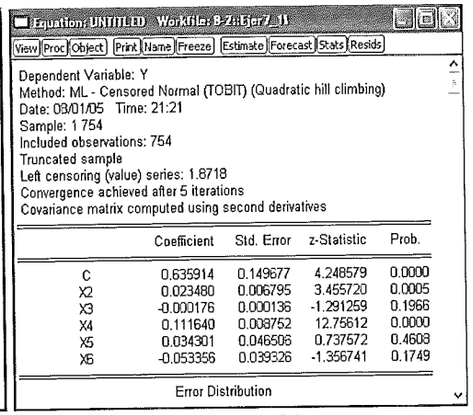


Figura 5-80

El efecto marginal en el modelo Tobit truncado de cada una de las variables se obtiene mediante la expresión (aplicada en los valores medios):

$$\frac{\partial E(Y_i \mid X_i; Y_i > a)}{\partial X_k} = \beta_k - \beta_k[\{\gamma(\alpha_i)\}^2 - \alpha_i\gamma(\alpha_i)] = \beta_k[1-[\{\gamma(\alpha_i)\}^2 - \alpha_i\gamma(\alpha_i)]] = \beta_k[1-\delta(\alpha_i)]$$

Luego el efecto marginal de variable X_k respecto de la variable explicada equivale al coeficiente estimado β_k multiplicado por el factor de corrección $[1-\delta(\alpha_i)]$ siendo:

$$\delta(\alpha_i) = \{\gamma(\alpha_i)\}^2 - \alpha_i\gamma(\alpha_i) \qquad \gamma(\alpha_i) = \frac{\phi\left(\dfrac{a - X_i\beta}{\sigma}\right)}{1-\Phi\left(\dfrac{a - X_i\beta}{\sigma}\right)} \qquad \alpha_i = \frac{a - X_i\beta}{\sigma}$$

Tenemos:

$$\hat{\alpha}_i = \frac{a - X_i\hat{\beta}}{\hat{\sigma}} = \frac{1{,}8718 - 2{,}391155}{0{,}429388} = -1{,}209524$$

$$\gamma(\hat{\alpha}_i) = \frac{\phi\left(\dfrac{a - X_i\hat{\beta}}{\hat{\sigma}}\right)}{1-\Phi\left(\dfrac{a - X_i\hat{\beta}}{\hat{\sigma}}\right)} = \frac{\phi(-1{,}209524)}{1-\Phi(-1{,}209524)} = \frac{0{,}191970}{1-0{,}1131} = 0{,}216450$$

$$\delta(\hat{\alpha}_i) = \{\gamma(\hat{\alpha}_i)\}^2 - \hat{\alpha}_i\gamma(\hat{\alpha}_i) = 0{,}21645^2 - (-1{,}209524)0{,}21645 = 0{,}308652$$

Los efectos marginales para las variables X_2, X_3 y X_4 sobre el salario se calcularán como sigue:

$$\frac{\partial E(Y_i \mid X_i; Y_i > a)}{\partial X_{2i}} = \hat{\beta}_2[1-\delta(\hat{\alpha}_i)] = 0{,}02348(1-0{,}308652) = 0{,}01623285$$

$$\frac{\partial E(Y_i \mid X_i; Y_i > a)}{\partial X_{3i}} = \hat{\beta}_3[1-\delta(\hat{\alpha}_i)] = 0{,}000176(1-0{,}308652) = -0{,}00012167$$

$$\frac{\partial E(Y_i \mid X_i; Y_i > a)}{\partial X_{4i}} = \hat{\beta}_4[1-\delta(\hat{\alpha}_i)] = 0{,}11640(1-0{,}308652) = 0{,}077182$$

A continuación consideramos el archivo *casadas.wf1* que contiene datos de 753 mujeres casadas en 1975, de las cuales 428 participaron en el mercado laboral en algún momento del año (*inlf*=1) y 325 no participaron (*inlf*=0). Se considera que la participación en el mercado de trabajo (*inlf*) depende de otras fuentes de ingresos (*nwifeinc*), de los años de educación (*educ*), de los años de experiencia laboral (*exper* y *exper*^{2}), de la edad (*age*), del número de hijos menores de 6 años (*kidslt6*) y del número de hijos entre 6 y 18 años (*kidsg6*).

Basándose en la información anterior, mediante un modelo Tobit truncado estimaremos la ecuación salarial que hace depender el logaritmo del salario de las mujeres (*wage*) de su nivel educativo, de su experiencia laboral y del cuadrado de esta última variable usando el método en dos etapas de Heckman.

Al igual que en el caso anterior estamos ante un ***modelo Tobit con truncamiento incidental*** porque la participación en el mercado laboral depende de la oferta de trabajo.

En una primera etapa estimamos un modelo Probit de *Inlf* frente a la constante y todas las variables explicativas. El modelo es:

$$Inlf = \Phi(\beta_0 + \beta_1\, nwifeinc + \beta_2\, educ + \beta_3\, exper + \beta_4\, exper^2 + \beta_5\, age + \beta_6\, kidslt6 + \beta_7\, kidsg6)$$

Para ello se elige *Quick* → *Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *BINARY-Binary chice (logit, probit, extreme value)* en el campo *Method*, se elige *Probit* en el campo *Binary estimation method* para realizar el ajuste mediante el modelo Probit (Figura 5-81) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 5-82 con buenas significatividades individuales para los parámetros estimados (salvo la constante y *kidsge6*). La significatividad conjunta es muy alta porque el p-valor del estadístico de la razón de verosimilitud es muy pequeño. El Pseudo R^2 de McFadden no se acerca demasiado a la unidad (0,22). Los valores de los criterios de información (Akaike, Schwarz y Hannan-Quinn) son adecuados porque son bajos y muy parecidos.

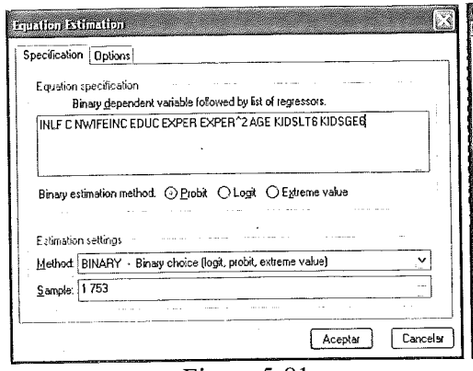


Figura 5-81

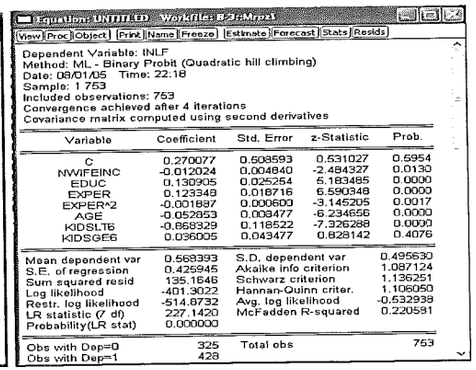


Figura 5-82

A continuación calculamos el inverso del ratio de Mills a través de los residuos generalizados del modelo Probit recién ajustado. Para ello, en la ventana del Probit se elige *Procs* → *Make residual series* y se elige *Generalized* (Figura 5-83). Se obtiene la nueva variable *resid*01 = $\hat{\lambda}_i$ equivalente al ratio de Mills (Figura 5-84).

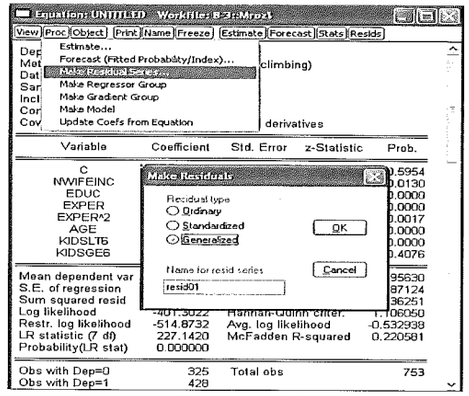


Figura 5-83

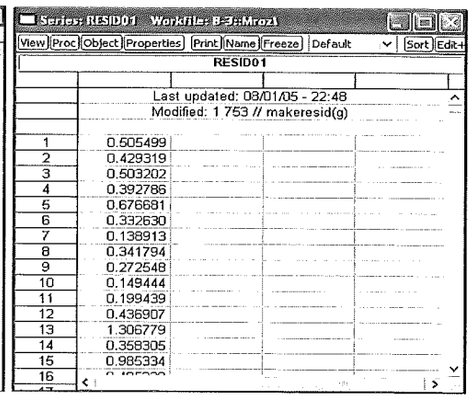


Figura 5-84

En una segunda etapa, estimamos por MCO la ecuación salarial incluyendo como regresor adicional el residuo generalizado del Probit. El modelo a estimar será:

$$Log(wage) = \beta_0 + \beta_1\, educ + \beta_2\, exper + \beta_3\, exper^2 + \rho\, \hat{\lambda}_i + e$$

Para ello se elige *Quick* → *Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *LS-Two stage Least Squares (NLS and ARMA)* en el campo *Method* para ajustar por mínimos cuadrados (Figura 5-85) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 5-86.

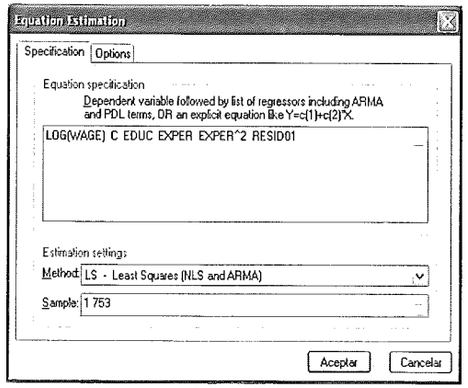


Figura 5-85

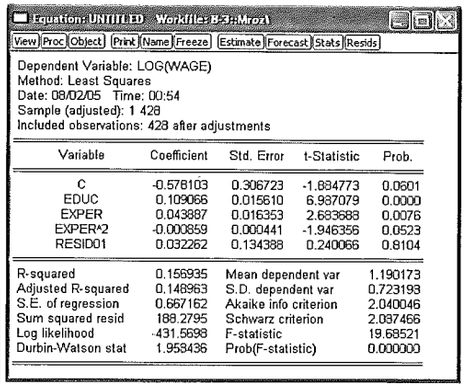


Figura 5-86

Se observa una muy buena significatividad de los coeficientes estimados, tanto individual como conjunta, buenos valores de los criterios de información y un estadístico de Durbin Watson muy cercano a 2, lo que indica que no existirán problemas de autocorrelación serial.

MODELOS DE VARIABLE DEPENDIENTE LIMITADA CON STATA: LOGIT Y PROBIT

STATA dispone del comando *logit*, que permite ajustar un modelo logístico binario. Como ejemplo partiremos del archivo de automóviles *auto.dta* y estimaremos un modelo que explique la probabilidad de que un coche sea extranjero en función de su peso y su consumo. El modelo será entonces:

$$Pr(foreign = 1) = F(\beta_0 + \beta_1\, weight + \beta_2\, mpg) \quad F = \text{distribución logística}$$

y se estima con STATA mediante la sintaxis siguiente:

```
. use http://www.stata-press.com/data/r9/auto
(1978 Automobile Data)

. logit foreign weight mpg

Iteration 0:   log likelihood =  -45.03321
Iteration 1:   log likelihood = -29.898968
Iteration 2:   log likelihood = -27.495771
Iteration 3:   log likelihood = -27.184006
Iteration 4:   log likelihood = -27.175166
Iteration 5:   log likelihood = -27.175156
Logistic regression                               Number of obs   =         74
                                                  LR chi2(2)      =      35.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -27.175156                       Pseudo R2       =     0.3966

------------------------------------------------------------------------------
     foreign |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0039067   .0010116    -3.86   0.000    -.0058894    -.001924
         mpg |  -.1685869   .0919174    -1.83   0.067    -.3487418     .011568
       _cons |   13.70837   4.518707     3.03   0.002     4.851864    22.56487
```

Los *odds ratio* se obtienen mediante el procedimiento *logistic*.

```
logistic foreign weight mpg

Logistic regression                               Number of obs   =         74
                                                  LR chi2(2)      =      35.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -27.175156                       Pseudo R2       =     0.3966

------------------------------------------------------------------------------
     foreign | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .9961009   .0010077    -3.86   0.000     .9941279    .9980779
         mpg |   .8448578   .0776572    -1.83   0.067     .7055753    1.011635
------------------------------------------------------------------------------
```


STATA dispone del comando *probit*, que permite ajustar un modelo *probit* binario. Como ejemplo partiremos del archivo de automóviles *auto.dta* y estimaremos un modelo que explique la probabilidad de que un coche sea extranjero en función de su peso y su consumo. El modelo será entonces:

$$Pr(foreign = 1) = \Phi(\beta_0 + \beta_1\, weight + \beta_2\, mpg) \quad \Phi = \text{distribución normal } (0,1)$$

y se estima con STATA mediante la sintaxis siguiente:

```
. probit foreign weight mpg

Iteration 0:   log likelihood =  -45.03321
Iteration 1:   log likelihood = -29.244141
Iteration 2:   log likelihood = -27.041557
Iteration 3:   log likelihood =  -26.84658
Iteration 4:   log likelihood = -26.844189
Iteration 5:   log likelihood = -26.844189

Probit regression                                 Number of obs   =         74
                                                  LR chi2(2)      =      36.38
                                                  Prob > chi2     =     0.0000
Log likelihood = -26.844189                       Pseudo R2       =     0.4039

------------------------------------------------------------------------------
     foreign |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0023355   .0005661    -4.13   0.000     -.003445   -.0012261
         mpg |  -.1039503   .0515689    -2.02   0.044    -.2050235   -.0028772
       _cons |   8.275464   2.554142     3.24   0.001     3.269438    13.28149
------------------------------------------------------------------------------
```

STATA dispone del comando *mlogit* para ajustar ***modelos logísticos multinomiales***. Como ejemplo se ajusta un modelo que explica la probabilidad de disponer de un seguro con tres modalidades alternativas (*insure*) que son prepago, indemnización y no seguro, en función de la raza del asegurado (*nonwhite*) considerando blancos y no blancos como razas posibles. La sintaxis de STATA es la siguiente:

```
. use http://www.stata-press.com/data/r9/sysdsn3
(Health insurance data)

. mlogit insure nonwhite

Iteration 0:   log likelihood = -556.59502
Iteration 1:   log likelihood = -551.78935
Iteration 2:   log likelihood = -551.78348
Iteration 3:   log likelihood = -551.78348

Multinomial logistic regression                   Number of obs   =        616
                                                  LR chi2(2)      =       9.62
                                                  Prob > chi2     =     0.0081
Log likelihood = -551.78348                       Pseudo R2       =     0.0086
```

```
------------------------------------------------------------------------------
      insure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Prepaid      |
    nonwhite |   .6608212   .2157321     3.06   0.002     .2379942    1.083648
       _cons |  -.1879149   .0937644    -2.00   0.045    -.3716896   -.0041401
-------------+----------------------------------------------------------------
Uninsure     |
    nonwhite |   .3779585    .407589     0.93   0.354    -.4209012    1.176818
       _cons |  -1.941934   .1782185   -10.90   0.000    -2.291236   -1.592632
------------------------------------------------------------------------------
(insure==Indemnity is the base outcome)
```

Según esta salida, la probabilidad de prepago para los blancos (*nonwhite*=0) se modeliza como sigue:

$$\Pr(insure = prepaid) = \frac{e^{-0,188}}{1 - e^{-0,188} + e^{-1,942}} = 0,42$$

Del mismo modo, para los no blancos la probabilidad de prepago será:

$$\Pr(insure = prepaid) = \frac{e^{-0,188+0,661}}{1 - e^{-0,188+0,661} + e^{-1,942+0,378}} = 0,57$$

STATA dispone del comando *mprobit* para ajustar ***modelos probit multinomiales***. Para el ejemplo anterior, el modelo *probit* multinomial se ajusta mediante la sintaxis siguiente:

```
. mprobit insure nonwhite

Iteration 0:   log likelihood = -552.66429
Iteration 1:   log likelihood = -551.78515
Iteration 2:   log likelihood = -551.78348
Iteration 3:   log likelihood = -551.78348

Multinomial probit regression                     Number of obs   =        616
                                                  Wald chi2(2)    =       9.52
Log likelihood = -551.78348                       Prob > chi2     =     0.0086

------------------------------------------------------------------------------
      insure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Prepaid      |
    nonwhite |   .5669085   .1837347     3.09   0.002     .2067951    .9270219
       _cons |  -.1616883   .0805726    -2.01   0.045    -.3196077   -.0037688
-------------+----------------------------------------------------------------
Uninsure     |
    nonwhite |   .3119427   .2616306     1.19   0.233    -.2008438    .8247291
       _cons |  -1.392729   .1145066   -12.16   0.000    -1.617157     -1.1683
------------------------------------------------------------------------------
(insure=Indemnity is the base outcome)
```


MODELOS TOBIT CENSURADO Y TRUNCADO CON STATA. MÉTODO DE HECKMAN Y RATIO DE MILLS

STATA utiliza el comando *tobit* para ajustar ***modelos con censura***. Como ejemplo ajustamos un modelo que explique el consumo de los automóviles en función de su peso (dividido por 1000) censurando los datos en el sentido de que consumos menores o iguales a 17 se sustituyen por el valor 17 (censura por la izquierda). El ajuste se haría ahora mediante un modelo *tobit* como sigue:

```
. use http://www.stata-press.com/data/r9/auto
(1978 Automobile Data)

. generate wgt = weight/1000

. replace mpg=17 if mpg<=17
(14 real changes made)

. tobit mpg wgt, ll

Tobit regression                                  Number of obs   =         74
                                                  LR chi2(1)      =      72.85
                                                  Prob > chi2     =     0.0000
Log likelihood = -164.25438                       Pseudo R2       =     0.1815

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wgt |  -6.87305   .7002559    -9.82   0.000    -8.268658   -5.477442
       _cons |  41.49856    2.05838    20.16   0.000     37.39621     45.6009
-------------+----------------------------------------------------------------
      /sigma |  3.845701   .3663309                      3.115605    4.575797
------------------------------------------------------------------------------
  Obs. summary:         18  left-censored observations at mpg<=17
                        56     uncensored observations
                         0 right-censored observations
```

A continuación ajustamos el modelo anterior pero suponiendo que no hay valores de *mpg* superiores a 24, es decir, en presencia de ***censura por la derecha***. La sintaxis para el modelo *tobit* correspondiente podría ser la siguiente:

```
. use http://www.stata-press.com/data/r9/auto, clear
(1978 Automobile Data)

. generate wgt = weight/1000

. tobit mpg wgt, ul(24)

Tobit regression                                  Number of obs   =         74
                                                  LR chi2(1)      =      90.72
                                                  Prob > chi2     =     0.0000
Log likelihood =  -129.8279                       Pseudo R2       =     0.2589
```

```
------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wgt |  -5.080645     .43493   -11.68   0.000    -5.947459   -4.213831
       _cons |   36.08037   1.432056    25.19   0.000     33.22628    38.93445
-------------+----------------------------------------------------------------
      /sigma |   2.385357   .2444604                      1.898148    2.872566
------------------------------------------------------------------------------
  Obs. summary:          0  left-censored observations
                        51     uncensored observations
                        23 right-censored observations at mpg>=24
```

Si ahora consideramos el ***modelo censurado por la izquierda*** y por la derecha, el modelo *tobit* correspondiente se ajustaría como sigue:

```
. tobit mpg wgt, ll(17) ul(24)

Tobit regression                                  Number of obs   =         74
                                                  LR chi2(1)      =      77.60
                                                  Prob > chi2     =     0.0000
Log likelihood = -104.25976                       Pseudo R2       =     0.2712

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wgt |  -5.764448   .7245417    -7.96   0.000    -7.208457   -4.320438
       _cons |   38.07469   2.255917    16.88   0.000     33.57865    42.57072
-------------+----------------------------------------------------------------
      /sigma |   2.886337   .3952143                      2.098676    3.673998
------------------------------------------------------------------------------
  Obs. summary:         18  left-censored observations at mpg<=17
                        33     uncensored observations
                        23 right-censored observations at mpg>=24
```

STATA ajusta los ***modelos con sesgo de selección*** mediante el comando *heckman*. Como ejemplo se ajusta un modelo que explica el salario en función del nivel educativo y la edad en presencia de sesgo de selección relativo a las variables estado civil, hijos, nivel educativo y edad en el sentido de que el salario sólo se observa si una combinación lineal de estas últimas variables es positiva.

```
. clear

. use http://www.stata-press.com/data/r9/womenwk

. heckman wage educ age, select(married children educ age)

Iteration 0:   log likelihood = -5178.7009
Iteration 1:   log likelihood = -5178.3049
Iteration 2:   log likelihood = -5178.3045

Heckman selection model                           Number of obs     =       2000
(regression model with sample selection)          Censored obs      =        657
                                                  Uncensored obs    =       1343

                                                  Wald chi2(2)      =     508.44
Log likelihood = -5178.304                        Prob > chi2       =     0.0000
```



```
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wage         |
   education |   .9899537   .0532565    18.59   0.000     .8855729    1.094334
         age |   .2131294   .0206031    10.34   0.000     .1727481    .2535108
       _cons |   .4857752   1.077037     0.45   0.652    -1.625179     2.59673
-------------+----------------------------------------------------------------
select       |
     married |   .4451721   .0673954     6.61   0.000     .3130794    .5772647
    children |   .4387068   .0277828    15.79   0.000     .3842534    .4931601
   education |   .0557318   .0107349     5.19   0.000     .0346917    .0767718
         age |   .0365098   .0041533     8.79   0.000     .0283694    .0446502
       _cons |  -2.491015   .1893402   -13.16   0.000    -2.862115   -2.119915
-------------+----------------------------------------------------------------
     /athrho |   .8742086   .1014225     8.62   0.000     .6754241    1.072993
    /lnsigma |   1.792559    .027598    64.95   0.000     1.738468     1.84665
-------------+----------------------------------------------------------------
         rho |   .7035061   .0512264                      .5885365    .7905862
       sigma |   6.004797   .1657202                       5.68862    6.338548
      lambda |   4.224412   .3992265                      3.441942    5.006881
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =    61.20   Prob > chi2 = 0.0000
------------------------------------------------------------------------------
```

MODELO DE POISSON CON STATA

STATA permite trabajar con el modelo de datos de recuento de Poisson mediante el comando *poisson*. A continuación se presenta un ejemplo.

```
. use http://www.stata-press.com/data/r9/airline

. poisson injuries XYZowned

Iteration 0:   log likelihood = -31.507676
Iteration 1:   log likelihood = -31.507676

Poisson regression                                Number of obs   =          9
                                                  LR chi2(1)      =       0.80
                                                  Prob > chi2     =     0.3698
Log likelihood = -31.507676                       Pseudo R2       =     0.0126

------------------------------------------------------------------------------
    injuries |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    XYZowned |  -.2451225   .2780192    -0.88   0.378    -.7900301    .2997852
       _cons |   2.036882    .147442    13.81   0.000     1.747901    2.325863
------------------------------------------------------------------------------
```

Ejercicio 5-1. En un estudio sobre el mercado inmobiliario se dispone de información sobre el nivel de renta de las familias (RENTA), sobre el número de hijos de la familiaa (HIJOS) sobre una variable dicotómica (SUB) que toma el valor 1 si la vivienda está subvencionada y 0 en otro caso y sobre la variable discreta CASA referida a los metros cuadrados de las viviendas que están a disposición de los compradores cuyo valor depende de la superficie de la siguiente forma:

$$CASA = \begin{cases} 0 \text{ si la vivienda tiene menos de } 80\ \text{m}^2 \\ 1 \text{ si la vivienda tiene entre } 81\ \text{m}^2 \text{ y } 100\ \text{m}^2 \\ 2 \text{ si la vivienda tiene entre } 101\ \text{m}^2 \text{ y } 120\ \text{m}^2 \\ 3 \text{ si la vivienda tiene más de } 120\ \text{m}^2 \end{cases}$$

Se trata de analizar qué tipo de vivienda compran las familias según sus características (renta e hijos) y las de la vivienda (subvencionada o no) utilizando un modelo de respuestra múltiple $Y_i^* = F(X_i\beta) + u_i$ y suponiendo que la función de distribución F pueda ser una normal (0,1) o una logística ¿Qué método es mejor?

Como la variable dependiente tiene valores que se pueden ordenar de menor a mayor según su superficie, se puede especificar un modelo de respuesta múltiple ordenado del tipo $Y_i^* = F(X_i\beta) + u_i$ con:

$$Y_i = \begin{cases} 0 & si\ Y_i^* \le c_1 \\ 1 & si\ c_1 \le Y_i^* \le c_2 \\ 2 & si\ c_2 \le Y_i^* \le c_3 \\ 3 & si\ c_3 \le Y_i^* \end{cases}$$

Los distintos modelos a utilizar se presentan al considerar la función de distribución F como una normal (modelo Probit Multidimensional) o una logística (modelo Logit Multidimensional).

Para realizar la ***estimación del modelo Probit Multinomial ordenado*** se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *ORDERED-Ordered choice* en el campo *Method,* se elige *Normal* en el campo *Error ditribution* para realizar el ajuste mediante el modelo Probit Multinomial (Figura 5-87) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 5-88 con buenas significatividades individuales para los parámetros estimados (salvo quizá *SUB* con una significatividad del 85% aproximadamente). La significatividad conjunta es muy alta porque el p-valor del estadístico de la razón de verosimilitud es muy pequeño. El Pseudo R^2 se acerca bastante al 60%. Los valores de los criterios de información (Akaike, Schwarz y Hannan-Quinn) son adecuados.

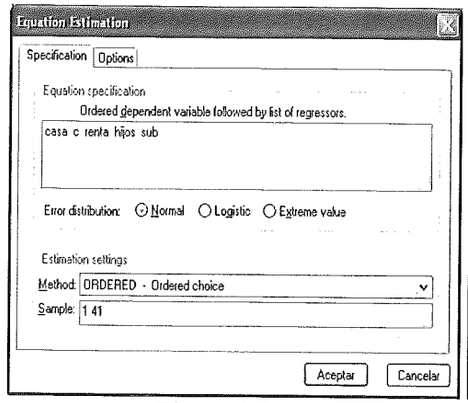


Figura 5-87

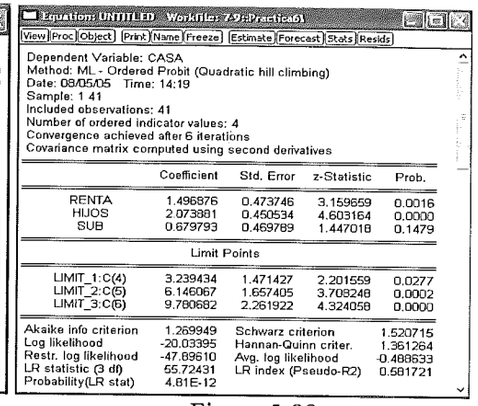


Figura 5-88

Para realizar la ***estimación del modelo Logit Multinomial ordenado*** se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *ORDERED-Ordered choice* en el campo *Method,* se elige *Logistic* en el campo *Error distribution* para realizar el ajuste mediante el modelo Logit Multinomial (Figura 5-89) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 5-90 con buenas significatividades individuales para los parámetros estimados (salvo quizá *SUB* con una significatividad del 85% aproximadamente). La significatividad conjunta es muy alta porque el p-valor del estadístico de la razón de verosimilitud es muy pequeño. El Pseudo R^2 se acerca bastante al 60%. Los valores de los criterios de información (Akaike, Schwarz y Hannan-Quinn) son adecuados. Se observa que lo resultados de los estadísticos de la bondad de ajuste son muy similares al caso del modelo Probit.

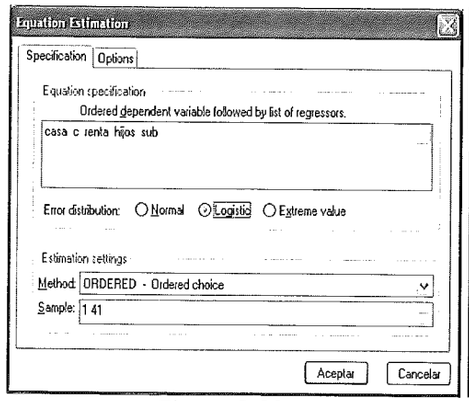


Figura 5-89

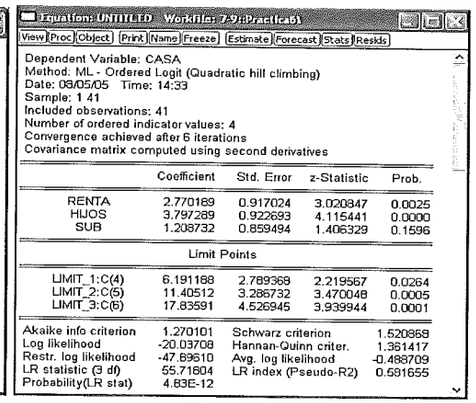


Figura 5-90

El método de estimación más eficiente será aquel que presente menores valores de los criterios de información Akaike, Schwarz y Hannan-Quinn y mayor valor de la función de verosimilitud. Recopilando tenemos:

	Modelo Probit	Modelo Logit
Log Likelihood	**-20,03395**	**-20,03708**
Akaike	**1,269949**	**1,270101**
Schwarz	**1,520715**	**1,520868**
Hannan-Quin	**1,361264**	**1,361417**

El método que presenta menores valores de los criterios de información y mayor valor de la función de verosimilitud es el modelo Probit. Por tanto, el mejor ajuste lo ofrece el modelo Probit Multinomial ordenado.

Ejercicio 5-2. En un estudio sobre la dotación de televisores en los hogares españoles se ha encuestado a 940 familias preguntando sobre el número de televisores en el hogar (TV), los ingresos anuales en millones de unidades monetarias (IF) y el nivel de instrucción del cabeza de familia (NI). La variable TV puede tomar los valores 1, 2 o 3 según las familias tengan 1, 2 y 3 o más televisores disponibles. El nivel de instrucción del cabeza de familia se ha codificado a través de cuatro niveles (Básica, Bachillerato superior, Universitaria media y Universitaria superior) con los valores respectivos enteros de 1 a 4. Con los datos especificados contenidos en el archivo 7-8.wf1, se pide ajustar un modelo logit que explique la probabilidad de que un hogar tenga un determinado número de televisores en función del resto de las variables anteriormente definidas.

Para una familia cuyo cabeza de familia tenga un nivel de instrucción básico (NI=1) y unos imgresos anuales de tres millones de unidades monetarias (IF=3) hallar la probabilidad de que tenga un sólo televisor. Calcular también la probabilidad de que esta familia tenga 2 televisores. Asimismo, calcular la probabilidad de que esta familia tenga tres o más televisores.

Calcular el efecto marginal de tener dos televisores para una familia cuyos ingresos anuales sean de cuatro millones de unidades monetarias y donde el nivel de instrucción del cabeza e familia sea básico respecto al caso en que este nivel sea de Bachillerato superior. Los datos se encuentran en el archjivo 5-2.wf1

Como la variable respuesta tiene más de dos categorías que pueden ordenarse, estamos ante el modelo logit multinomial ordenado:

$$TV_i = \Lambda(X\beta_i) = \Lambda(\beta_0 + \beta_1 IF_i + \beta_2 NI_i) + u_i = \frac{1}{1+e^{\beta_0+\beta_1 IF_i+\beta_2 NI_i}} + u_i$$

Para realizar la ***estimación del modelo Logit Multinomial ordenado*** se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *ORDERED-Ordered choice* en el campo *Method,* se elige *Logistic* en el campo *Error ditribution* para realizar el ajuste mediante el modelo Logit Multinomial (Figura 5-91) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 5-92 con buenas significatividades individuales para los parámetros estimados. La significatividad conjunta es muy alta porque el p-valor del estadístico de la razón de verosimilitud es muy pequeño. El Pseudo R^2 se acerca bastante a la unidad (0,88). Los valores de los criterios de información (Akaike, Schwarz y Hannan-Quinn) son adecuados.

Para una familia cuyo cabeza de familia tenga un nivel de instrucción básico (NI=1) y unos ingresos anuales de tres millones de unidades monetarias (IF=3), la probabilidad de que tenga un sólo televisor vedrá dada por:

$$Prob(TV=1) = \Lambda(\hat{\alpha}_1 - X_i\hat{\beta}) = \frac{1}{1+e^{-(\alpha_1-(\beta_1 IF+\beta_2 NI))}} = \frac{1}{1+e^{-(8,01055-(1,686776*3+0,820545*1)}} = 0,893754$$

Para la familia anterior la probabilidad de que tenga dos televisores será:

$$Prob(TV=2) = \Lambda(\hat{\alpha}_2 - X_i\hat{\beta}) - \Lambda(\hat{\alpha}_1 - X_i\hat{\beta}) = \frac{1}{1+e^{-(24,75738-(1,686776*3+0,820545*1)}} -$$

$$-\frac{1}{1+e^{-(8,01055-(1,686776*3+0,820545*1)}} = 0,999999 - 0,893754 = 0,106245$$

Para la familia anterior la probabilidad de que tenga tres o más televisores será:

$$Prob(TV=3) = 1 - \Lambda(\hat{\alpha}_2 - X_i\hat{\beta}) = 1 - \frac{1}{1+e^{-(24,75738-(1,686776*3+0,820545*1)}} =$$

$$= 1 - 0,999999 = 0,000001$$

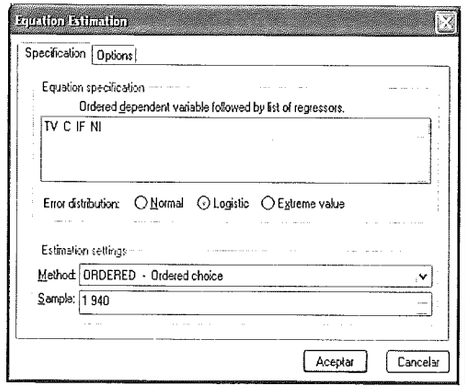


Figura 5-91

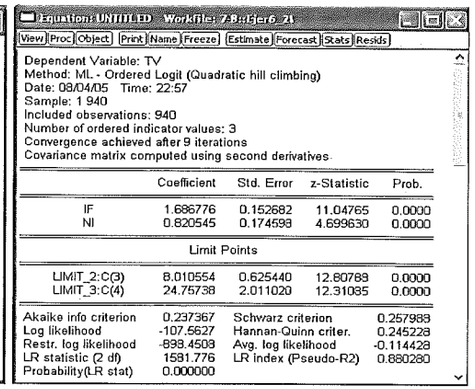


Figura 5-92

Para calcular el efecto marginal de tener dos televisores para una familia cuyos ingresos anuales sean de cuatro millones de unidades monetarias y donde el nivel de instrucción del cabeza de familia sea básico respecto al caso en que este nivel sea de Bachillerato superior, se deberá calcular la probabilidad para cada una de las situaciones y obtener la diferencia entre ellas. Tenemos:

$$Prob(TV=2\mid IF=4,NI=1)=\Lambda(\hat{\alpha}_2-X_i\hat{\beta})-\Lambda(\hat{\alpha}_1-X_i\hat{\beta})=\frac{1}{1+e^{-(24,75738-(1,686776*4+0,820545*1)}}-$$

$$-\frac{1}{1+e^{-(8,01055-(1,686776*4+0,820545*1)}}=0,999999-0,608950=0,391049$$

$$Prob(TV=2\mid IF=4,NI=2)=\Lambda(\hat{\alpha}_2-X_i\hat{\beta})-\Lambda(\hat{\alpha}_1-X_i\hat{\beta})=\frac{1}{1+e^{-(24,75738-(1,686776*4+0,820545*2)}}-$$

$$-\frac{1}{1+e^{-(8,01055-(1,686776*4+0,820545*2)}}=0,999999-0,406695=0,593304$$

El efecto marginal pedido se obtiene como la diferencia entre las dos probabilidades anteriores:

$$Prob(TV=2\mid IF=4,NI=2)-Prob(TV=2\mid IF=4,NI=1)=$$
$$=0,593304-0,391049=0,202255$$

Ejercicio 5-3. El archivo 5-3.wf1 contiene datos de 1445 presos sobre los meses que pasan hasta que un interno es detenido otra vez después de haber sido puesto en libertad (durat). Se considera que log(durat) depende de si el preso ha participado o no en programas de empleo en prisión (workprg), del número de condenas previas (priors), del número de meses pasados en prisión (tserved), de si consume o no alcohol (alcohol), de si consume o no drogas (drugs), de si el delito cometido es grave o no (felon), de si es negro o no (black), de si esta casado o no (married), de su nivel educativo (educ) y de su edad (age). Se sabe también que de los 1445 presos hay 893 que no fueron arrestados mientras se les siguió. Se dispone también de una variable ficticia que nos indica si cada observación está censurada o no (cens).

Basándose en la información anterior, mediante un modelo Tobit estimar e interpretar la relación que liga a la variable log(durat) con el resto de las variables explicativas, interpretando el resultado.

Como existen presos que no fueron arrestados después de haber sido puestos en libertad la última vez, estamos ante un modelo Tobit censurado por la derecha, estando definida la censura por la variable *cens* que juega el papel de indicador de censura.

Realizaremos la ***estimación del modelo Tobit censurado por la derecha por máxima verosimilitud***. Para ello se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *CENSORED-Censored or truncated data (Tobit)* en el campo *Method,* se sitúa la variable *cens* en el campo *Right* y se señala *Zero/One indicador of censoring* porque el modelo Tobit está censurado a la derecha siendo el indicador de censura la variable *cens*, se elige *Normal* porque consideramos el modelo con distribución normal (Figura 5-93) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 5-94 con buenas significatividades individuales para los parámetros estimados (salvo *workprg* y *educ*). Los valores de los criterios de información no son demasiado elevados y son bastante coincidentes entre ellos.

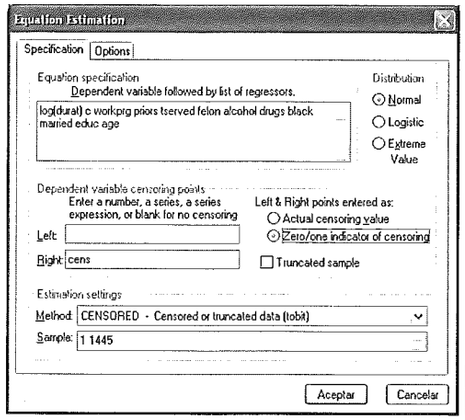


Figura 5-93

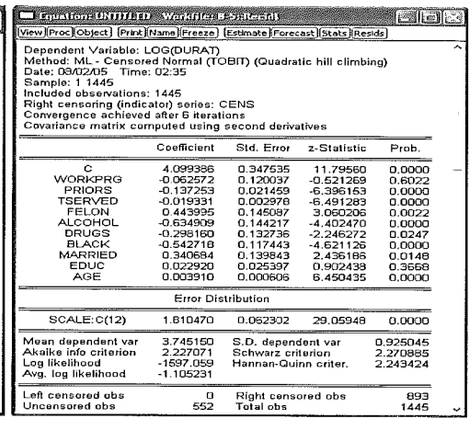


Figura 5-94

Se observa que las variables *priors* y *tserved* tienen un efecto negativo. Un preso con una condena más reducirá el tiempo que transcurre hasta que vuelve a ser detenido en un 14%. Un mes más en prisión reduce el tiempo que transcurre hasta ser detenido en un 1,9%. Un individuo que consume drogas (*drugs* = 1) tarda en volver a ser arrestado aproximadamente un 29,8% menos $(e^{-0,298}-1)\times 100=-25,78$.

Un individuo que es negro tarda en volver a ser arrestado aproximadamente un 54,3% menos que un blanco $(e^{-0,543}-1)\times 100=-41,8$.

Se observa que si estimamos el modelo por MCO sin tener en cuenta que 893 de las 1445 observaciones están censuradas, obtenemos resultados diferentes. Para ello se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *LS-Two stage Least Squares (NLS and ARMA)* en el campo *Method* para ajustar por mínimos cuadrados (Figura 5-95) y se hace clic en *Aceptar*.

Se obtienen los resultados de la Figura 5-96 con buenas significatividades individuales para los parámetros estimados (salvo *workprg* y *educ*) y conjunta. Los valores de los criterios de información no son demasiado elevados y son bastante coincidentes entre ellos. El estadístico de Durbin Watson está muy cercano a 2, lo que indica que no existirán problemas de autocorrelación serial. Se observa un ajuste que no coincide con el del modelo Tobit censurado. Aunque las estimaciones de los parámetros no difieren en demasía, uno de ellos cambia de signo (*workprg*), que precisamente es una de las variables individualmente no significativa.

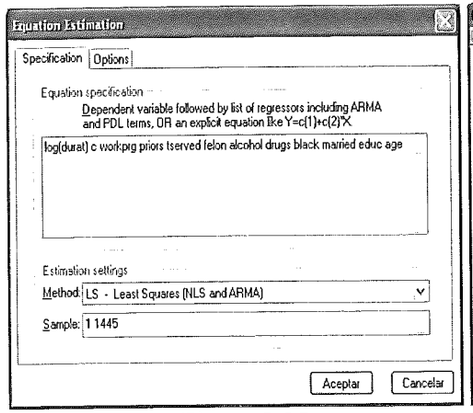


Figura 5-95

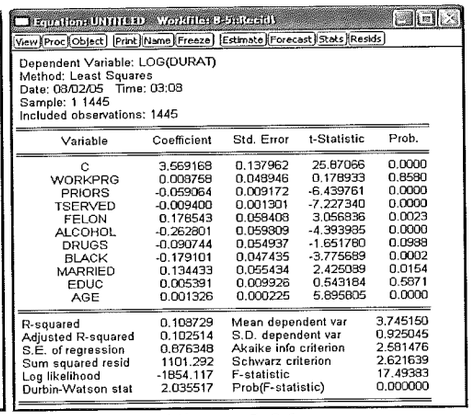


Figura 5-96

Ejercicio 5-4. Se trata de estudiar los efectos de dos calmantes para el dolor de cabeza. Se dividen los pacientes en dos grupos de modo que cada grupo recibe un tipo diferente de calmante (TIPO) registrándose para cada paciente el tiempo que tarda en remitir el dolor de cabeza después de haberle sido suministrado el calmante (MINUT). Como puede haber pacientes que al final del período de observación sigan con dolor de cabeza, estamos ante un caso de datos censurados por la derecha. La variable censura (CENS) valdrá 1 para datos censurados y cero para datos no censurados. Los datos se recogen ene el archivo 5-4.wf1.

Ajustar un modelo adecuado que explique el tiempo de remisión del dolor de cabeza en función del grupo de pacientes, lo que permitirá comparar ambos calmantes. Ajustar los modelos suponiendo distribición normal, logística y de valor extremo. Comparar los tres modelos.

Como puede haber pacientes que al final del período de observación sigan con dolor de cabeza, estamos ante un caso de un modelo Tobit censurado por la derecha, estando definida la censura por la variable CENS que juega el papel de indicador de censura.

Realizaremos la ***estimación del modelo Tobit censurado por la derecha por máxima verosimilitud***. Para ello se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification*, se elige *CENSORED-Censored or truncated data (Tobit)* en el campo *Method,* se sitúa la variable *cens* en el campo *Right* y se señala *Zero/One indicador of censoring* porque el modelo Tobit está censurado a la derecha siendo el indicador de censura la variable *CENS*, se elige *Normal* porque consideramos el modelo con distribución normal (Figura 5-97) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 5-98 con buenas significatividades individuales para los parámetros estimados. Los valores de los criterios de información son aceptables y coincidentes entre ellos.

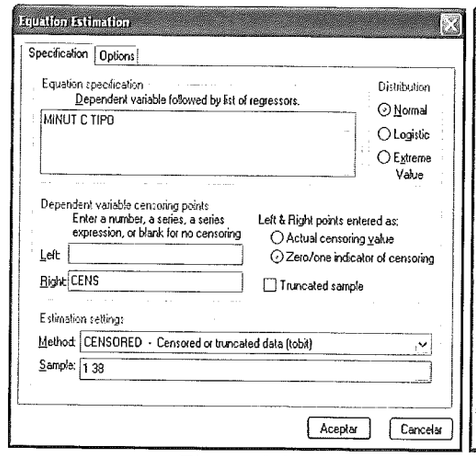


Figura 5-97

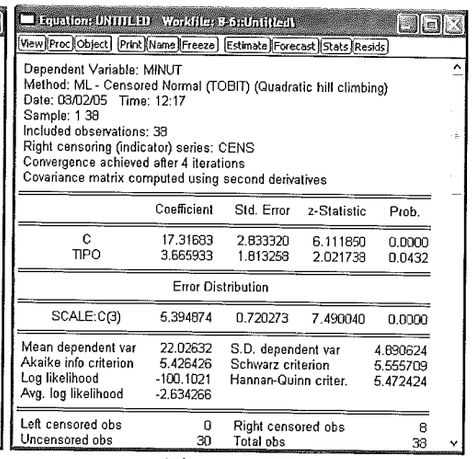


Figura 5-98

Se observa que para un paciente del grupo 1 (*TIPO* = 1), el tiempo de remisión del dolor de cabeza es de 17,317 + 3,666*1 = 20,983 minutos, mientras que para uno del grupo 2 (TIPO = 2) es de 17,317 + 3,666*2 = 24,649 minutos. Se observa que el calmante suministrado al grupo 1 es más efectivo porque el dolor remite en menos tiempo que con el calmante suministrado al grupo 2.

Para ajustar el modelo Tobit censurado por la derecha suponiendo distribución logística, señalamos *Logistic* en el campo *Distribution* de la pantalla *Equation Estimation* (Figura 5-99). Si se hace clic en *Aceptar* se obtienen los resultados de la Figura 5-100 con buenas significatividades individuales para los parámetros estimados. Los valores de los criterios de información son aceptables y coincidentes entre ellos. Se observa además, que las diferencias con el modelo de distribución *Normal* son muy pequeñas.

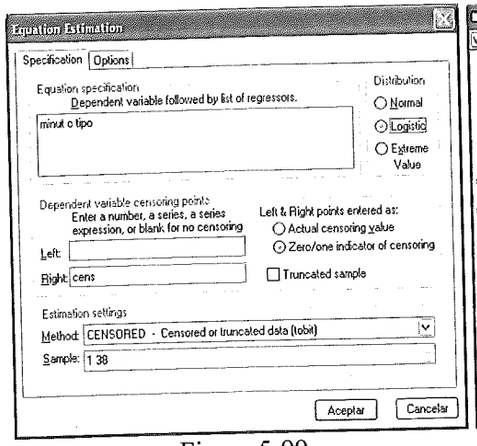


Figura 5-99

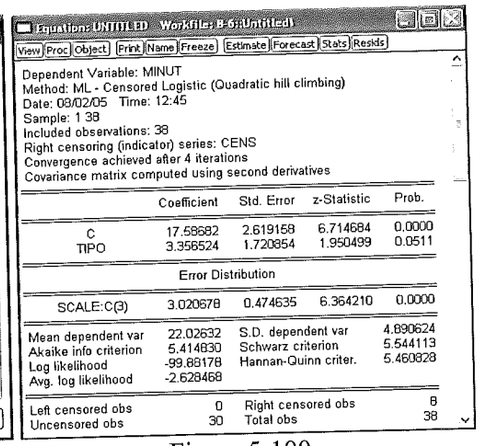


Figura 5-100

Para ajustar el modelo Tobit censurado por la derecha suponiendo distribución logística, señalamos *Extreme value* en el campo *Distribution* de la pantalla *Equation Estimation* (Figura 5-101). Si se hace clic en *Aceptar* se obtienen los resultados de la Figura 5-102 con buenas significatividades individuales para los parámetros estimados. Los valores de los criterios de información son aceptables y coincidentes entre ellos. Se observa además, que las diferencias con los modelos de distribución *Normal* y *Logit* son pequeñas.

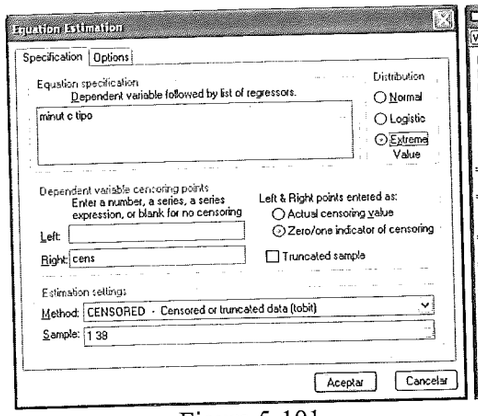


Figura 5-101

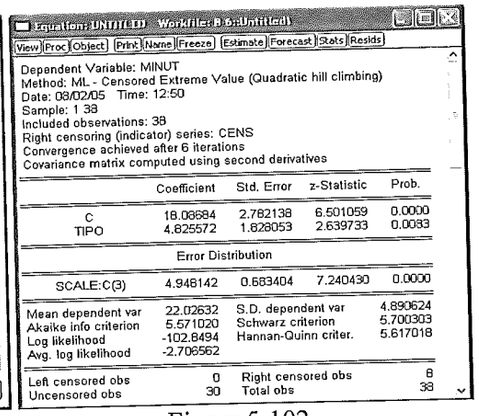


Figura 5-102

Realizada la estimación de los modelos Tobit para las tres distribuciones, realizaremos una comparación entre ellos, para quedarnos con el más adecuado. El modelo más eficiente será aquel que presente menores valores de los criterios de información de Schwarz y Hannan-Quinn y mayor valor de la función de verosimilitud. Recopilando tenemos:

	Normal	Logit	Valor extremo
Log Likelihood	-100,1021	-99,88178	-102,8494
Schwarz	5,555709	5,544113	5,700303
Hannan-Quin	5,472424	5,460828	5,617018

El método que presenta menores valores de los criterios de información y mayor valor de la función de verosimilitud es el modelo con distribución logística con muy poquita diferencia respecto del modelo con distribución normal. El modelo con distribución de Weibull (valor extremo) es el menos adecuado en este caso. Por lo tanto, el mejor ajuste lo ofrece el modelo Tobit censurado por la derecha con distribución logística.

Ejercicio 5-5. Consideramos el número de lingotes r no listos para poner en circulación de entre n contrastados, para un número de combinaciones de tiempos de exposición al calor y de tiempos de remojo. Con estos datos se trata de ajustar un modelo logit que estima la probabilidad de que un lingote no pueda ser puesto en circulación.

La sintaxis SAS adecuada podría ser la siguiente:

```
data lingotes;
     input calor remojo r n @@;
     datalines;

   7 1.0 0 10  14 1.0 0 31  27 1.0 1 56  51 1.0 3 13
   7 1.7 0 17  14 1.7 0 43  27 1.7 4 44  51 1.7 0  1
   7 2.2 0  7  14 2.2 2 33  27 2.2 0 21  51 2.2 0  1
   7 2.8 0 12  14 2.8 0 31  27 2.8 1 22  51 4.0 0  1
   7 4.0 0  9  14 4.0 0 19  27 4.0 1 16
  ;
proc logistic data=lingotes;
     model r/n=calor remojo;
run;
```

La salida es la siguiente:

```
                    The LOGISTIC Procedure

                       Model Information
Data Set                          WORK.LINGOTES
Response Variable (Events)        r
Response Variable (Trials)        n
Number of Observations            19
Link Function                     Logit
Optimization Technique            Fisher's scoring
                     Response Profile

        Ordered      Binary          Total
          Value      Outcome      Frequency

              1      Event               12
              2      Nonevent           375
```

```
                    Model Convergence Status

          Convergence criterion (GCONV=1E-8) satisfied.
                      Model Fit Statistics
                                              Intercept
                              Intercept          and
            Criterion           Only        Covariates

            AIC                108.988         101.346
            SC                 112.947         113.221
            -2 Log L           106.988          95.346

            Testing Global Null Hypothesis: BETA=0

      Test                  Chi-Square       DF     Pr > ChiSq

      Likelihood Ratio        11.6428         2         0.0030
      Score                   15.1091         2         0.0005
      Wald                    13.0315         2         0.0015

                      The LOGISTIC Procedure
              Analysis of Maximum Likelihood Estimates

                                Standard
Parameter     DF    Estimate      Error     Chi-Square     Pr > ChiSq

Intercept      1     -5.5592     1.1197        24.6503         <.0001
calor          1      0.0820     0.0237        11.9454         0.0005
remojo         1      0.0568     0.3312         0.0294         0.8639

                     Odds Ratio Estimates
                       Point          95% Wald
            Effect   Estimate     Confidence Limits

            calor       1.085       1.036       1.137
            remojo      1.058       0.553       2.026

      Association of Predicted Probabilities and Observed Responses

            Percent Concordant     64.4    Somers' D     0.460
            Percent Discordant     18.4    Gamma         0.555
            Percent Tied           17.2    Tau-a         0.028
            Pairs                  4500    c             0.730
```

El ajuste presenta el criterio de información de Akakie y el criterio de Schwarz (utilizados para elegir aquel modelo que presente menores cantidades para estos valores) y –2LogL. Los contrastes de razón de verosimilitud (*Likelihood Ratio*), *score* y *Wald* aceptan la significatividad conjunta al 95% de las variables explicativas, ya que sus p-valores son menores que 0,05. La tabla de estimaciones de máxima verosimilitud presenta los parámetros estimados, sus errores estándar y los resultados del test de Wald para la significatividad individual de los parámetros estimados (el único parámetro no significativo al 95% es *remojo* cuyo p-valor es mayor que 0,05).

La función logística estimada será:

$$p = \frac{1}{1+e^{-(-5,5592+0,082 calor+0,0568 humedad)}}$$

Según este ajuste, la probabilidad de que un lingote con tiempo de exposición al calor = 7 y tiempo en remojo = 1 no pueda ser puesto en circualción, vendrá dada por:

$$p = \frac{1}{1+e^{-(-5,5592+0,082*7+0,0568*1)}} = \frac{1}{1+e^{-(4,9284)}} = 0,0072$$

En la última tabla de la salida se presentan medidas de asociación para contrastar la capacidad predictiva del modelo.

Ejercicio 5-6. Ssupongamos que queremos contrastar el efecto de una droga en 12 niveles de dosificación y que para ello dividimos aleatoriamente 180 sujetos en 12 grupos (uno por cada nivel de dosificación) de 15 sujetos cada uno. Para cada sujeto registramos la presencia o ausencia de respuesta positiva a la droga y resumimos los datos contando el número de sujetos que responden positivamente a la droga en cada grupo de dosis (variable respuesta). La variable dosis representará la cantidad de droga suministrada, la variable numero representará el número de sujetos en cada grupo (vale 15 para todas las observaciones). Se trata de modelizar la probabilidad de respuesta positiva como una función de la dosis suministradamediante un modelo probit.

La sintaxis SAS adecuada podría ser la siguiente:

```
data estudio;
      input Dosis Respuesta;
      Numero = 15;
      Observada=Respuesta/Numero;
      datalines;
   0     3
   1.1   4
   1.3   4
   2.0   3
   2.2   5
   2.8   4
   3.7   5
   3.9   9
   4.4   8
   4.8  11
   5.9  12
   6.8  13
  ;
   run;
proc probit data=estudio log10 optc;
      model respuesta/numero=dosis;
      output out=nuevo p=p_hat;
   run;
```

La salida es la siguiente:

```
                    Probit Procedure
                    Model Information
          Data Set                          WORK.ESTUDIO
          Events Variable                      Respuesta
          Trials Variable                         Numero
```

```
                Number of Observations                    12
                Number of Events                          81
                Number of Trials                         180
                Number of Events In Control Group          3
                Number of Trials In Control Group         15
                Name of Distribution                  NORMAL
                Log Likelihood              -104.3945783

 Algorithm converged.

                      Analysis of Parameter Estimates
                                Standard
Variable          DF  Estimate    Error Chi-Square Pr > ChiSq Label

Intercept          1  -4.14385  1.34149     9.5419     0.0020 Intercept
Log10(Dosis)       1   6.23076  1.89958    10.7588     0.0010
_C_                    0.24088  0.05226                        Lower threshold

              Probit Model in Terms of Tolerance Distribution
                                MU        SIGMA
                        0.66506312   0.16049411

           Estimated Covariance Matrix for Tolerance Parameters
                          MU            SIGMA            _C_
MU              0.001158          -0.000493        0.000954
SIGMA          -0.000493           0.002394       -0.000999
_C_             0.000954          -0.000999        0.002731
```

Según estos datos, el modelo $Pi=F(x_i\beta)+e_i$ puede ajustarse mediante la ecuación $Pi=F(-4{,}1439+6{,}2308 log10(dosis))$. Además la distribución de tolerancia tiene media 0,665 y desviación típica 0,16 y presenta la matriz de covarianzas que aparece al final de la salida. La ecuación del modelo probit ajustado es:

$$p=(2\pi)^{-1/2}\int_{-\infty}^{-4{,}1439+6{,}2308\times\log10(\text{dosis})} e^{-\frac{t^2}{2}}dt$$

Ejercicio 5-7. Como segundo ejemplo consideramos la variable Estimulo que indica el nivel de estímulo aplicado a un determinado número de sujetos N de entre los cuales una cantidad responde positivamente al estímulo (variable Respuesta). Se trata de estimar la probabilidad de que un individuo responda positivamente al estímulo utilizando el modelo logit y el modelo probit.

Se ajustará un modelo del tipo:

$$p = Pr(\ Respuesta) = F\ (\ b_0 + b_1\ \times log_{10}(\ Estimulo))$$

donde F es la función de distribución de una normal (0,1) para el caso del modelo probit, y F es la función logística en el caso del modelo logit. La sintaxis SAS podría ser la siguiente:

```
data a;
      infile cards eof=eof;
      input estimulo N Respuesta;
      Observado= Respuesta/N;
      output;
      return;
```



```
eof: do Dose=0.5 to 7.5 by 0.25;
        output;
     end;
   datalines;
 1 10 1
 2 12 2
 3 10 4
 4 10 5
 5 12 8
 6 10 8
 7 10 10
;
 proc probit log10;
    model Respuesta/N=Estimulo / lackfit inversecl itprint;
    model Respuesta/N=Estimulo / d=logistic inversecl;
    output out=B p=Prob std=std xbeta=xbeta;
    title 'Salida del procedimiento Probit'; run;
```

La salida con el ajuste probit es la siguiente:

```
                        Salida del procedimiento Probit

                            Probit Procedure
                           Model Information
                   Data Set                          WORK.A
                   Events Variable                Respuesta
                   Trials Variable                        N
                   Number of Observations                 7
                   Number of Events                      38
                   Number of Trials                      74
                   Missing Values                        29
                   Name of Distribution              NORMAL
                   Log Likelihood             -37.28038802

Algorithm converged.
                            Goodness-of-Fit Tests
                Statistic                    Value      DF     Pr > ChiSq
                Pearson Chi-Square          3.6497       5         0.6009
                L.R.    Chi-Square          4.6381       5         0.4616

                          Response-Covariate Profile
                        Response Levels              2
                        Number of Covariate Values   7

Since the chi-square is small (p > 0.1000), fiducial limits will be calculated using a t value of
1.96.
                          Analysis of Parameter Estimates
                                     Standard
          Variable          DF  Estimate    Error Chi-Square Pr > ChiSq Label

          Intercept          1  -1.81270  0.44934    16.2743     <.0001 Intercept
          Log10(estimulo)    1   3.41812  0.74555    21.0196     <.0001

                   Probit Model in Terms of Tolerance Distribution
                                MU        SIGMA
                        0.53032254   0.29255866

                        Estimated Covariance Matrix
                         for Tolerance Parameters
                                MU            SIGMA
                   MU         0.002418       -0.000409
                   SIGMA     -0.000409        0.004072
```

La salida con el ajuste logit es la siguiente:

```
                    Probit Procedure
                   Model Information
Data Set                              WORK.A
Events Variable                     Respuesta
Trials Variable                           N
Number of Observations                    7
Number of Events                         38
Number of Trials                         74
Missing Values                           29
Name of Distribution                 LOGISTIC
Log Likelihood                 -37.11065336

Algorithm converged.
                Analysis of Parameter Estimates
                              Standard
Variable          DF   Estimate     Error Chi-Square Pr > ChiSq Label
Intercept          1   -3.22464   0.88606    13.2447     0.0003 Intercept
Log10(estimulo)    1    5.97018   1.44917    16.9721     <.0001

          Probit Model in Terms of Tolerance Distribution
                          MU          SIGMA
                  0.54012513     0.16749914

                 Estimated Covariance Matrix
                  for Tolerance Parameters
                              MU             SIGMA
MU                      0.002378          -0.000381
SIGMA                  -0.000381           0.001653
```

En ambos modelos, logit y probit, los parámetros estimados son significativos al 99% (p-valores menores que 0,01) y el ajuste global también es bueno (los contrastes de la bondad de ajuste tienen p-valores mayores que 0,01). El modelo probit ajustado es:

$$p = Pr(\,Respuesta) = F\,(\,-1{,}81 + 3{,}41 \times log_{10}(\,Estimulo))=$$

$$=(2\pi)^{-1/2}\int_{-\infty}^{-1{,}81+3{,}41\times log10(\,Estimulo)} e^{-\frac{t^2}{2}}\,dt$$

El modelo logit ajustado resulta ser el siguiente:

$$p = Pr(\,Respuesta) = F\,(\,-3{,}22 + 5{,}97 \times log_{10}(\,Estimulo))=$$

$$=\frac{e^{-3{,}22+5{,}97\times log10(\,Estimulo)}}{1+e^{-3{,}22+5{,}97\times log10(\,Estimulo)}}$$

CAPÍTULO 6

ANÁLISIS UNIVARIANTE DE SERIES TEMPORALES. MODELOS ARIMA, INTERVENCIÓN Y FUNCIÓN DE TRANSFERENCIA

SERIES TEMPORALES

Dentro de las estructuras de datos más importantes, típicas en el trabajo econométrico aplicado, tenemos los datos de series temporales. Un conjunto de datos de series temporales consiste en observaciones sobre una variable o distintas variables a lo largo del tiempo. Ejemplos típicos de datos de series temporales son el producto interior bruto, la oferta monetaria, los índices de precios al consumo, las tasas anuales de homicidios o las cifras de venta de automóviles. Dado que los acontecimientos pasados pueden tener influencia sobre acontecimientos futuros, y los efectos retardados en el comportamiento de los individuos son frecuentes en ciencias sociales, el tiempo es un parámetro importante en los conjuntos de series temporales.

Los datos de series temporales suelen utilizarse más en el análisis macroeconómico, en contraposición a los datos de corte transversal, que se utilizan sobre todo en el análisis microeconómico. Las series temporales suelen ser más difíciles de analizar que los datos de corte transversal debido a que casi nunca podemos suponer que las observaciones económicas son temporalmente independientes. La mayoría de las series temporales, ya sean económicas o no, están relacionadas (a menudo fuertemente relacionadas) con su historia reciente. Por ejemplo, nuestro conocimiento sobre el producto nacional bruto del trimestre pasado nos dice bastante del nivel de PIB que podemos esperar para el trimestre en curso ya que el PIB tiende a permanecer estable de un trimestre a otro. Otra característica importante de los datos de series temporales es la periodicidad con la que se recogen (semanal, mensual, trimestral, etc.) con el hecho adicional de que muchas series temporales semanales, mensuales o trimestrales muestran una característica estacional marcada que puede ser un factor importante en la metodología del análisis de dichas series.

DESCOMPOSICIÓN CLÁSICA DE UNA SERIE TEMPORAL

Una serie temporal es una sucesión de valores en el tiempo. Designaremos una serie temporal por Y_{ik}, donde el índice i toma los valores 1, 2, ..., N representando por ejemplo años, y el índice k toma los valores 1, 2, ..., m representando por ejemplo meses ($m = 12$) o trimestres ($m = 4$), o cualquier otra fracción de año. La teoría clásica considera una serie de tiempo formada por cuatro componentes teóricas: tendencia, variaciones estacionales, variaciones cíclicas y variaciones residuales (Figura 6-1).

La *tendencia* viene dada por el movimiento general a largo plazo de la serie. Designaremos a la tendencia por T_{ik}. Las *variaciones estacionales* son oscilaciones que se producen con un periodo igual o inferior a un año, y que se reproducen de manera reconocible en los diferentes años. Designaremos a las variaciones estacionales por E_{ik}. Las *variaciones cíclicas* son oscilaciones que se producen con un periodo superior al año, y que se deben principalmente a la alternancia de etapas largas (ciclos) en las que se repite el comportamiento de la serie. Designaremos a las variaciones cíclicas por C_{ik}. Las *variaciones residuales o irregulares* son movimientos en la serie que no muestran un carácter periódico reconocible y que son originados por fenómenos singulares que afectan a la variable en estudio de manera casual y no permanente. Designaremos a las variaciones residuales por R_{ik}.

Las componentes teóricas de una serie temporal pueden combinarse de diferentes formas, dando lugar a distintos esquemas de formación de la serie. El *esquema aditivo* supone que $Y_{ik} = T_{ik} + C_{ik} + E_{ik} + R_{ik}$; el *esquema multiplicativo* supone que $Y_{ik} = T_{ik}.C_{ik}.E_{ik}.R_{ik}$; el *esquema mixto* supone que $Y_{ik} = T_{ik}.C_{ik}.E_{ik} + R_{ik}$. Un supuesto fundamental del análisis clásico es la independencia de las variaciones residuales respecto de las demás componentes.

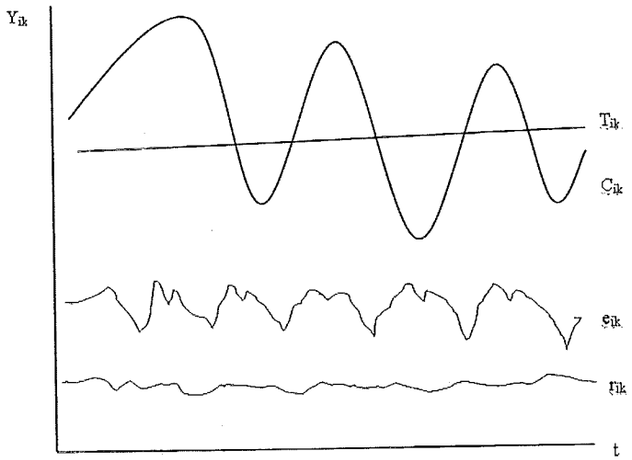


Figura 6-1

Tendencia de una serie temporal: ajuste analítico, medias móviles y diferencias

Centrándonos ya en el análisis de la tendencia, designaremos a la serie temporal por Z_t, dependiendo sólo del índice t (periodo de tiempo principal), ya que de lo que se trata es de aislar el movimiento a largo plazo de la serie (no usamos el otro subíndice porque, al hacer un estudio a largo plazo, no es relevante la subdivisión de cada periodo principal en subperiodos).

Método de ajuste analítico

Para hallar la tendencia de una serie temporal mediante ajuste analítico, realizamos un ajuste por regresión de los valores de la serie a una función del tiempo que sea sencilla, y que recoja de manera satisfactoria la marcha general del fenómeno representado por la serie temporal. Es común considerar entre otras las funciones de ajuste $Z(t) = a + bt$ (lineal), $Z(t) = a + bt + ct^2$ (cuadrática), y $Z(t) = Exp(a+bt)$ (exponencial). No obstante pueden realizarse ajustes a tendencias de todo tipo (logarítmicas, semilogarítmicas, polinómicas, potenciales, hiperbólicas...)

Tendencia lineal. Una línea de tendencia lineal es una línea recta $Z(t) = a + bt$ que se ajusta correctamente a los datos. Una línea de tendencia lineal normalmente muestra que algo aumenta o disminuye a un ritmo constante.

Tendencia logarítmica. Una línea de tendencia logarítmica $Z(t) = log(a+bt)$ es una línea curva muy útil cuando el índice de cambios de los datos aumenta o disminuye rápidamente y, después, se estabiliza.

Tendencia polinómica. Una línea de tendencia polinómica $Z(t) = a + bt + ct^2 + ... + ct^n$ es una línea curva que se utiliza cuando los datos fluctúan según la ecuación de un polinomio. Es útil, por ejemplo, para analizar las pérdidas y ganancias de un conjunto de datos grande. El orden del polinomio se puede determinar mediante el número de fluctuaciones en los datos, o en función del número de máximos y mínimos que aparecen en la curva. Una línea de tendencia polinómica de *orden 2* suele tener sólo un máximo o un mínimo. Una de *orden 3* normalmente tiene uno o dos máximos o mínimos. El *orden 4* tiene más de tres.

Tendencia potencial. Una línea de tendencia de potencia es una línea curva $Z(t) = at^b$ que se utiliza con conjuntos de datos que comparan medidas que aumentan a un ritmo concreto; por ejemplo, la aceleración de un automóvil de carreras a intervalos de un segundo. No es posible crear una línea de tendencia de potencia si los datos contienen valores cero o negativos.

Tendencia exponencial. Una línea de tendencia exponencial es una línea curva $Z(t) = Exp(a+bt)$ que es muy útil cuando los valores de los datos aumentan o disminuyen a intervalos cada vez mayores. No es posible crear una línea de tendencia exponencial si los datos contienen valores cero o negativos.

Tendencia de media móvil. Una línea de tendencia de media móvil atenúa las fluctuaciones en los datos para mostrar con mayor claridad la trama o la tendencia. Una media móvil utiliza un número concreto de puntos de datos (establecido por la opción *Periodo*), hace un promedio de los mismos, y utiliza el valor del promedio como punto de la línea; por ejemplo, si el valor de *Periodo* está establecido en 2, el promedio de los dos primeros puntos de datos se utiliza como el primer punto en la línea de tendencia de media móvil. El promedio de los puntos de los datos segundo y tercero se utiliza como el segundo punto de la línea de tendencia, etc.

Método de las medias móviles

El método de las medias móviles de orden p analiza la tendencia de una serie temporal a partir del resumen de los datos iniciales mediante determinadas medias de los mismos elaboradas de la siguiente forma:

Si p es impar se forman medias relativas a los instantes $(p+1)/2$, $(p+3)/2$, $(p+5)/2$, ... (que serán valores enteros porque p es impar). La serie de medias es la siguiente:

$$\bar{y}_{\frac{p+1}{2}} = \frac{y_1 + y_2 + \cdots y_p}{p}, \quad \bar{y}_{\frac{p+3}{2}} = \frac{y_2 + y_3 + \cdots y_{p+1}}{p}, \quad \bar{y}_{\frac{p+5}{2}} = \frac{y_3 + y_3 + \cdots y_{p+2}}{p}, \cdots$$

Si p es par se forman medias relativas a los instantes $(p+1)/2$, $(p+3)/2$, $(p+5)/2$, ... (que no serán valores enteros porque p es par). A continuación se hallan nuevas medias móviles entre cada dos medias móviles originales consecutivas, que serán ahora relativas a los instantes $(p+2)/2$, $(p+4)/2$, $(p+6)/2$, ... (que ya serán valores enteros porque p es par). La serie de medias móviles es la siguiente:

$$\bar{y}_{\frac{p+2}{2}} = \frac{\bar{y}_{\frac{p+1}{2}} + \bar{y}_{\frac{p+3}{2}}}{p}, \quad \bar{y}_{\frac{p+4}{2}} = \frac{\bar{y}_{\frac{p+3}{2}} + \bar{y}_{\frac{p+5}{2}}}{p}, \cdots$$

Una vez obtenida la serie de medias móviles, la tendencia será la línea que las une.

Método de las diferencias

Consiste en derivar de la serie original y_t una nueva serie z_t obtenida como la diferencia entre el valor de la variable en el momento actual y el valor en el momento inmediatamente anterior $z_t = y_t - y_{t-1} = \nabla y_t$. Se puede comprobar si z_t crece o decrece a largo plazo, o si oscila alrededor de un mismo valor. En este segundo caso la serie ya no tendría tendencia, pero en el primero habría que seguir calculando una nueva serie de diferencias w_t definida como $w_t = z_t - z_{t-1} = \nabla z_t = \nabla\nabla y_t = \nabla^2 y_t$ y así sucesivamente hasta encontrar un serie aleatoria sin tendencia.

Variaciones estacionales: medias móviles, diferencias estacionales y variables ficticias

Ya sabemos que las variaciones estacionales son oscilaciones que se producen con un periodo igual o inferior a un año, y que se reproducen de manera reconocible en los diferentes años.

El motivo principal que induce a estudiar la componente estacional es que en la inmensa mayoría de las series económicas dicha componente provoca una distorsión de su verdadero movimiento. Para eliminar estas distorsiones y captar el movimiento real de la serie, es necesario eliminar las oscilaciones estacionales desestacionalizando la serie.

La desestacionalización es una tarea no trivial que ha dado lugar a multitud de estudios y algoritmos, entre los que destacan los programas X11 y X12 del *Bureau of the Census* de Estados Unidos. A nivel trivial, existen varios métodos de desestacionalización. Los más sencillos son el método de la tendencia, el método de las medias móviles y el método de las diferencias estacionales, a los que intentaremos aproximarnos aquí.

Cuando se representa una serie temporal mediante y_t, se suponen todas las observaciones ordenadas una detrás de otra tal y como se van produciendo ($t = 1, 2, \ldots, T$). Cuando representamos una serie temporal por y_{ik}, estamos considerando explícitamente el año i ($i = 1, 2, \ldots, N$) y la estación del año k ($k = 1, 2, \ldots, m$). Cuando la estación es el año, $m = 12$, y cuando es el trimestre, $m = 4$. Siempre se tiene que $T = Nm$.

Método de desestacionalización de la tendencia o método de las relaciones de medias mensuales respecto a la tendencia

El *método de desestacionalización de la tendencia* consta de los pasos siguientes:

- Ajustar una recta por mínimos cuadrados $\bar{y}_{i.} = a - bi$ a las medias anuales de los datos observados $\bar{y}_{i.} = \frac{1}{m}\sum_{k=1}^{m} y_{ik}$.
- Calcular las medias mensuales en los diferentes años $\bar{y}_{.k} = \frac{1}{N}\sum_{i=1}^{N} y_{ik}$ $k = 1, 2, \ldots, m$.
- Aislar la componente estacional obteniendo la serie de medias mensuales corregidas $\bar{y}'_{.k} = \bar{y}_{.k} - \frac{b(k-1)}{m}$.
- Calcular la media global corregida $\bar{y}' = \frac{\bar{y}'_{.1} + \bar{y}'_{.2} + \cdots + \bar{y}'_{.m}}{m}$.

- Si el esquema es multiplicativo, se calculan los índices de variación estacional $I_k = \frac{\bar{y}'_{.k}}{\bar{y}'}100$ y se desestacionaliza la serie dividiendo sus valores por los índices de variación estacional. La componente estacional es $E_{ik} = I_k/100$.
- Si el esquema es aditivo, la componente estacional del mes k es $E_{ik} = \bar{y}'_{.k} - \bar{y}'$.

Métodos de desestacionalización del índice estacional

Existen varios métodos de desestacionalización basados en el cálculo de índices estacionales. Aparte del explicado en el apartado anterior podemos citar un *método general de índice estacional* que consta de los siguientes pasos:

- Dada la serie cronológica por meses, estaciones, etc., en varios años, se halla la tendencia mediante el método de las medias móviles tomando un año de periodo.
- Se centran los valores así obtenidos en los instantes de tiempo originales.
- Se elimina la tendencia y la variación cíclica en ella incluida, dividiendo los datos de la serie original por los valores de la tendencia en cada instante del tiempo.
- Se eliminan las variaciones irregulares hallando las medias aritméticas de los valores observados en cada periodo de repetición anual.
- Sobre estos últimos valores se calculan los índices de variación estacional en forma de porcentajes.
- Se dividen los valores de la serie original por los índices de variación estacional correspondienes, obteniéndose la serie temporal desestacionalizada.

Método de desestacionalización de las medias móviles

El *método de desestacionalización de las medias móviles* consiste en obtener la componente extraestacional mediante un ajuste de la serie original por medias móviles de orden m para eliminar las variaciones estacionales. Un procedimiento de medias móviles simples para el ajuste estacional podría ser el siguiente:

Sea X_t (t = 1,2,...,n) una serie temporal estacional de periodo s (s = 4 para datos trimestrales y s = 12 para periodos mensuales). Una serie de medias móviles centrada de s puntos, X_t*, se obtiene a través de los siguientes pasos para s par:

Para medias móviles de s puntos $x^*_{t+0,5} = \frac{\sum_{j=-(s/2)+1}^{s/2} x_{t+j}}{s}$ $\left(t = \frac{s}{2}, \frac{s}{2}+1, \cdots, n-\frac{s}{2}\right)$

Para medias móviles centradas de s puntos $x^*_t = \frac{x^*_{t-0,5} + x^*_{t+0,5}}{2}$ $\left(t = \frac{s}{2}+1, \frac{s}{2}+2, \cdots, n-\frac{s}{2}\right)$

Método de las diferencias estacionales

El *método de desestacionalización de las diferencias estacionales* permite eliminar la mayor parte del efecto estacional de una serie, y consiste en obtener la serie de diferencias de orden m (periodo estacional), definida como $z_t = y_t - y_{t-m}$. De todos modos, es conveniente recordar que en cada diferenciación de orden m perdemos m observaciones de la serie original.

Variables ficticias en el análisis estacional

Son muy habituales las series de tiempo económicas basadas en información mensual o trimestral que presentan patrones estacionales. Suele ser útil eliminar la componente estacional de las series de tiempo con el fin de poderse concentrar en componentes más importantes como la tendencia. Ya sabemos que el proceso de eliminar la componente estacional de una serie de tiempo se conoce como *desestacionalización o ajuste estacional* y la serie obtenida se denomina serie desestacionalizada. Hay muchos métodos para desestacionalizar una serie temporal entre los que se encuentra el método de las variables ficticias dicotómicas.

Supongamos por ejemplo que tenemos una serie temporal Y_t con estacionalidad trimestral. Para desestacionalizarla consideramos el modelo:

$$Y_t = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t} + u_t$$

$$D_{it} = \begin{cases} 1 \text{ en el trimestre } i \\ 0 \text{ en el resto} \end{cases} \qquad i = 1, \cdots, 4$$

Se observa que en el modelo se omite la constante para evitar la colinealidad perfecta. Para que el efecto estacional esté presente, los parámetros estimados del modelo anterior han de ser significativamente distintos de cero individualmente.

Los residuos estimados de la regresión anterior $\hat{u}_t = Y_t - \hat{Y}_t$ serán los valores de la serie desestacionalizada.

Variaciones cíclicas

La componente cíclica de una serie temporal es la más difícil de detectar, pues a diferencia de la tendencia, que es un movimiento a largo plazo muy general, y de las variaciones estacionales, que tienen un periodo fijo, las variaciones cíclicas tienen un periodo no fácilmente identificable y en muchos casos incluso variable, siendo frecuente la existencia de ciclos que se superponen, lo que hace todavía más difícil su identificación.

En la práctica, para identificar el ciclo, suele eliminarse de la serie la tendencia y las variaciones estacionales, y después analizar la parte restante de la serie, que puede denotarse por $x_{ik} = c_{ik} + r_{ik}$. Incluso puede prescindirse del doble subíndice, ya que no existe variación estacional. De esta forma se intentarán detectar los ciclos en la serie x_t, mediante determinados métodos entre los que destaca el análisis armónico.

Una onda armónica tiene la ecuación $X_j = A\ Cos\ wj + B\ Sen\ wj$, o también puede expresarse como $X_j = R\ Cos(wj - \alpha)$. Ambas expresiones son equivalentes mediante las relaciones $R = (A^2 + B^2)^{1/2}$ y $\alpha = Arctan\ (B/A)$. R se denomina amplitud y proporciona el valor máximo de X_j. El valor $2\pi/\omega$ es el periodo o intervalo de tiempo necesario para que se produzca una oscilación completa; $\omega/2\pi$ es la frecuencia o número de oscilaciones que se producen entre dos momentos consecutivos de tiempo; ω es el ángulo expresado en radianes, y α es la fase que marca el valor de X_j en el origen. La Figura 6-2 aclara estos conceptos.

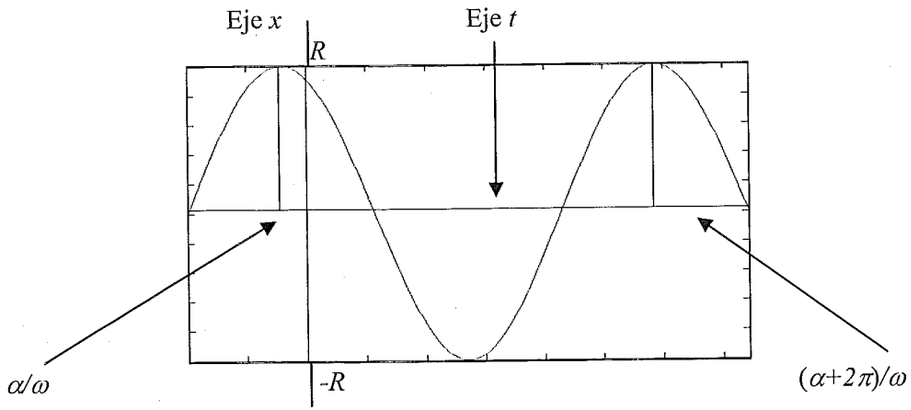


Figura 6-2

El trabajo fundamental en el análisis del ciclo es detectar en la serie original alguna función de tipo armónico o similar. Para detectar la existencia de un ciclo de orden p se suele formar el cuadro de la Figura 6-3.

1ª *oscilación*	x_1	x_2		x_p
2ª *oscilación*	x_{p+1}	x_{p+2}	...	x_{2p}
...	$\vdots$	$\vdots$		$\vdots$
q^a *oscilación*	$x_{(q-1)p+1}$	$x_{(q-1)p+2}$		x_{qp}
Medias	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_p$

Figura 6-3

La fila j-ésima de la tabla anterior recoge los p valores que forman la j-ésima oscilación, y el número de osilaciones q se obtiene dividiendo el número de observaciones de la serie x_t por el periodo p_j. La última fila de la tabla presenta las medias de los primeros elementos de cada oscilación, el valor medio de los segundos, etc.

La siguiente tarea es ajustar a los datos medios una expresión de la forma:

$$\bar{x}_j = A_0 + A\cos\frac{2\pi}{p}j + B\cos\frac{2\pi}{p}j \quad j = 1,2,\cdots,p$$

A_0 es una constante que se incluye para mejorar el ajuste, y $2\pi/p$ se incluye porque si el periodo es $2\pi/\omega$ entonces $\omega = 2\pi/p$.

El ajuste se realiza por mínimos cuadrados y las soluciones son las siguientes:

$$A_0 = \sum_{j=1}^{p}\frac{\bar{x}_j}{p}, \quad A = \frac{2}{p}\sum_{j=1}^{p}\bar{x}_j\cos\frac{2\pi j}{p}, \quad B = \frac{2}{p}\sum_{j=1}^{p}\bar{x}_j\operatorname{sen}\frac{2\pi j}{p}$$

Para distintos valores de p se obtienen distintas amplitudes $R(p)$ correspondientes a cada periodo. Los puntos $(p, R(p))$ forman el *periodograma*.

El periodograma transforma la serie temporal de su dominio natural, que es el tiempo, al dominio de las frecuencias (a los valores de la serie se le aplican transformadas de Fourier). Si no hay picos destacables en el periodograma no hay estacionalidad y cada pico destacable identifica un periodo que incluso puede ser un ciclo. A cada amplitud destacable le corresponde una frecuencia cuya inversa es el periodo estacional o cíclico. Luego el periodograma es un instrumento que identifica la longitud del periodo estacional y en su caso la del ciclo. Las amplitudes más fuertes (correspondientes a valores más bajos de las frecuencias p) suelen corresponder a ciclos, y las menos fuertes (correspondientes a valores no tan bajos de las frecuencias) suelen corresponder a estaciones.

También suele utilizarse el periodograma acumulativo que resulta de representar en el eje de abscisas las frecuencias y en el eje de ordenadas las amplitudes acumuladas. Para una serie aleatoria el periodograma acumulativo coincide con la diagonal del primer cuadrante. Desvíos bruscos de la diagonal provocan presencia de ciclos o estaciones para las respectivas frecuencias, que serán ciclos cuando las frecuencias sean bajas.

Existen diversos métodos para eliminar el ciclo en una serie temporal entre los que se encuentra el *filtro de Hodrick y Prescot*.

PREDICCIÓN Y SUAVIZADO DE SERIES TEMPORALES. MÉTODOS AUTOPROYECTIVOS DETERMINISTAS

Toda predicción es un intento de anticipar el futuro. En el contexto temporal, y tratándose de procedimientos cuantitativos, puede hablarse de dos clases de predicciones: condicionales e incondicionales. Las *predicciones condicionales* son las que se realizan mediante modelos causales. Por ejemplo, en un modelo de regresión que relaciona dos variables, una dependiente, Y, y otra independiente, X, las predicciones de Y están condicionadas a X, es decir, se predice Y dada X. Las *predicciones incondicionales* son las que se hacen mediante *métodos autoprotectivos* (el modelo de predicción sólo incluye valores actuales, pasados y futuros de la propia serie en estudio). Estos métodos pueden estar basados en dos enfoques alternativos: el determinista, o clásico, y el estocástico, o moderno (basado en la metodología de Box y Jenkins). El *enfoque determinista* es el que tratamos en este capítulo. El enfoque determinista es más adecuado cuando se dispone de un número limitado de observaciones, mientras que el enfoque estocástico es más adecuado cuando las series son de mayor tamaño.

Para cada tipo de predicciones (a corto, medio y largo plazo), existen determinados métodos más adecuados. Por ejemplo, el análisis de tendencias es un método para realizar predicciones a largo plazo, los modelos econométricos son adecuados para hacer predicciones a corto y medio plazo, y los métodos autoproyectivos son más adecuados para realizar predicciones a corto plazo. Precisamente, en las predicciones a corto plazo es conveniente tener presentes también las variaciones estacionales, lo mismo que en las predicciones a medio plazo es conveniente tener presente también la componente cíclica.

Suavizado por medias móviles

Los métodos autoprotectivos deterministas se utilizan para suavizar irregularidades y fluctuaciones de una serie temporal a fin de obtener la línea de suavizado como señal clara libre de variaciones estacionales y óptima para la predicción. Cuando no hay tendencia clara ni estacionalidad en la serie original, se utiliza el *suavizado por medias móviles*.

El método de medias móviles es un procedimiento mecánico para suavizar las irregularidades y las fluctuaciones de una serie temporal a fin de obtener la línea de tendencia. Dada la serie temporal X_t $t = 1, 2, ..., T$, se define la media móvil de orden, por ejemplo 5, como sigue:

$$T_s = (X_{s-2}+X_{s-1}+X_s+X_{s+1}+X_{s+2})/5 \qquad s = 3, 4, ..., T-2$$

La serie temporal T_s es una versión suavizada de la X_t.

Si se elige bien el orden de la media móvil, T_s no contiene la componente estacional, y será una representación correcta de las componentes a medio y largo plazo (ciclo-tendencia). El método es equivalente a ajustar una tendencia lineal a cada cinco puntos consecutivos de la serie inicial y tomar en cada ajuste solamente el punto central de la recta ajustada para alisar la serie original.

Suavizado lineal de Holt

El método de Holt, al igual que el de Brown, sirve para realizar predicciones bajo el supuesto de tendencia lineal, pero a diferencia de aquél, utiliza dos parámetros de alisado a y b, que toman valores constantes entre 0 y 1. Los valores predichos vienen dados por el modelo lineal: $F_t(l) = S_{t-1} + (b_{t-1})l$ $t > 2$, donde $S_t = aX_t + (1-a)[S_{t-1} + b_{t-1}]$ y $b_t = b[S_t - S_{t-1}] + (1-b)b_{t-1}$ .Los valores iniciales son $S_1 = x_1$ y $b_1 = x_2-x_1$

N = Número de observaciones
x_t = Observación t de la serie de tiempo en estudio
S_t = Observación t de la serie alisada
$F_t(l)$ = Predicción en el instante t a horizonte l
b_t = Valor del parámetro estimado del modelo en el instante t
a = Primera constante de alisado (relacionado con la componente aleatoria)
b = Segunda constante de alisado (relacionado con la tendencia)

Suavizado exponencial de Brown

Sea X_t el valor observado de la serie temporal (a la que vamos a aplicar el método suavizado) en el instante t. Sea $S_t(l)$ la predicción de X_t a horizonte l. S_t va a ser un suavizado de la serie X_t. El *modelo simple de alisado exponencial de Brown* obtiene predicciones de una serie temporal en función de las observaciones pasadas. Cada predicción se obtiene promediando los valores observados de la variable como sigue: $S_t(l) = aX_t + a(1-a)X^2_{t-1} + a(1-a)^2 X_{t-2} + ...$

Este procedimiento produce, efectivamente, un alisado de la serie X_t, ya que la nueva serie alisada, $S_t(l)$, al estar constituida por promedios (medias ponderadas) de valores de la serie primaria, presentará fluctuaciones más amortiguadas que X_t.

El valor de a ha de ser fijado entre cero y uno. Valores más pequeños de a alisan más los datos. Como regla práctica, si los datos presentan fuertes fluctuaciones o gran aleatoriedad, se deben usar valores pequeños de a. Las predicciones obtenidas mediante este procedimiento no cambian con el horizonte temporal, es decir, $S_t(1) = S_t(2) = ... = S_t(l) = ...$

En general podemos poner $S_t(l) = aX_t + (1 - a)S_{t-1}(l)$ para el *modelo simple de alisado exponencial de Brown*. En todos los métodos de predicción basados en el suavizado exponencial, se presenta el problema de la fijación de los valores iniciales.

Si el parámetro de alisamiento a está próximo a cero, el valor inicial fijado (S_0) influirá en el resultado durante muchos periodos de tiempo. Por el contrario, con valores de a próximos a la unidad, desaparecerá rápidamente la influencia del valor inicial adoptado, pero es muy posible que los datos presenten tendencias o estacionalidad, en cuyo caso usar este método de predicción no sería muy adecuado. Se toma como valor inicial $S_1 = X_1$.

También existe el *modelo exponencial de Brown con tendencia lineal*, extensión del modelo simple mediante $S'_t(l) = aX_t + (1 - a)S'_{t-1}(l)$ y $S''_t(l) = aS'_t + (1 - a)S'_{t-1}(l)$. En este caso suponemos la serie temporal generada por el esquema definido como $X_{t+i} = d + ei + u_{t+i}$ ($i = 0, 1, 2, 3, ...$). La predicción en el periodo t a horizonte l se obtendrá mediante $S_t(l) = a_t + b_t l$, donde a_t y b_t son estimaciones de d y e. Será necesario, una vez fijado a, dar un valor inicial a S_t' cuando $t = 1$. Este valor se establece siguiendo algún criterio ajeno al método. Se puede hacer, por ejemplo, $S_1' = X_1$. También se puede igualar S_1' a un promedio de los primeros valores de X. También habrá que dar un valor inicial a S_1'', que también puede ser X_1. También se toma $a_t = 2\,S'_t - S''_t$ y $b_t = a(S'_t - S''_t)/(1-a)$. Como valor fijado para a se suele tomar un número entre 0,1 y 0,3.

Además, existe el *modelo exponencial de Brown con tendencia cuadrática*, que es una extensión de los dos anteriores. En este caso suponemos que la serie original sigue una tendencia cuadrática de la forma: $X_{t+i} = d + ei + fi^2 + u_{t+i}$ ($i = 0, 1, 2, 3, ...$). Las ecuaciones para el modelo son $S'_t(l) = aX_t + (1 - a)S'_{t-1}(l)$, $S''_t(l) = aS'_t + (1 - a)S'_{t-1}(l)$ y $S'''_t(l) = aS''_t + (1 - a)S''_{t-1}(l)$. Las predicciones se obtendrán mediante la serie alisada $S_t(l) = p_t + (q_t)l + 1/2\,(r_t)l^2$, en donde, dado a, los parámetros $p_t = 3St' - 3St'' + St'''$, $q_t = a[(1-5a)S_t' - (10-8a)S_t'' + (4-3a)S_t''']/[2(1-a)^2]$ y $r_t = a^2(St' - 2St'' + St''')/(1-a)^2$ se estiman con los valores iniciales $S_1' = S_1'' = S_1''' = X_1$. El método con tendencia cuadrática sirve para predecir series con puntos de cambio de tendencia (*turning points*). Los métodos con tendencia lineal y simple no son válidos para este fin.

Suavizado estacional de Winters

Winters generalizó el método de Holt para tratar con datos que presenten variaciones estacionales. En este caso la fórmula de predicción es: $F_t(l) = (S_t + l(b_t))I_{t+1-L}$ donde L es el número de observaciones anuales. La estacionalidad se tiene en cuenta mediante el factor I_{t+1-L} (modelo multiplicativo). Las fórmulas de actualización son $S_t = aX_t/I_{t-L} + (1-a)[S_{t-1} + b_{t-1}]$, $bt = b(S_t - S_{t-1}) + (1-b)b_{t-1}$ y $I_t = cX_t/S_t + (1-c)I_{t-L}$.

Ahora es necesario conocer el valor de tres parámetros a, b y c. El primero está relacionado con la componente aleatoria, el segundo con la tendencia y el tercero con la componente estacional. La inicialización de los cálculos con este procedimiento requiere usar al menos L periodos para establecer los índices estacionales iniciales. Para estimar el factor de tendencia es conveniente utilizar datos referidos a $2L$ periodos (dos años consecutivos) en la forma siguiente:

$$b_1 = [(X_{L+1}-X_1)/L + (X_{L+2}-X_2)/L + ... + (X_{L+L}-X_L)]/L$$

PREDICCIONES INCONDICIONALES ESTOCÁSTICAS

En el capítulo anterior hemos estudiado las series temporales desde el punto de vista determinista o clásico. En este apartado vamos a ver el estudio de las series temporales desde el punto de vista estocástico o moderno, que utiliza métodos más complejos y su aplicación requiere series más largas. También sabemos del capítulo anterior que existen predicciones condicionales e incondicionales.

Las predicciones condicionales se realizan a través de modelos causales (se predicen valores futuros de la variable independiente de un modelo según los valores que tomen las variables independientes del modelo ajustado).

Las predicciones incondicionales se realizan mediante métodos autoprotectivos (se predicen valores futuros de una variable en función de valores pasados, actuales y futuros de la misma). Pero las predicciones incondicionales pueden tener un enfoque determinista o estocástico según la naturaleza del modelo utilizado.

El esquema siguiente ilustra la clasificación de las técnicas de predicción:

Predicciones
- *Condicionales → Se realizan mediante modelos causales (regresión, etc.)*
- *Incondicionales*
  - *Deterministas → Métodos autoproyectivos deterministas*
  - *Estocásticas → Métodos autoproyectivos estocásticos*

En este capítulo se estudian las predicciones incondicionales mediante métodos autoprotectivos con un enfoque estocástico a través de modelos ARIMA.

MODELOS ARIMA: PRIMEROS CONCEPTOS

Box y Jenkins son los autores de la modelización ARIMA. Un modelo ARIMA (*AutoRegresive Integrated Moving Average*) es un modelo estadístico autoproyectivo que permite predecir valores de una variable en función de sus valores pasados sin necesidad de ninguna otra información de variables auxiliares o relacionadas. Cada observación en un momento dado es modelada en función de valores previos suyos en el tiempo.

El nombre genérico ARIMA de estos modelos se deriva de sus tres componentes: *Autorregresivo* (AR), *Integrado*(I) de *Medias Móviles* (MA). El modelo ARIMA presenta una ecuación explícita que permite describir un valor como una función lineal de datos anteriores y errores debidos al azar. Puede incluir, además, un componente cíclico o estacional. El objetivo consiste en obtener un modelo adecuado, pero parsimonioso, es decir, el modelo ARIMA debe contener todos los elementos necesarios, pero los mínimos necesarios para describir el fenómeno en estudio. Box y Jenkins recomiendan como mínimo unas 50 observaciones en la serie temporal. Modelizar una serie temporal consiste en derivar un modelo ARIMA que se ajuste al conjunto de datos dado. Para ello es necesario estudiar características esenciales de las series como estacionalidad, estacionaridad, funciones de autocorrelación, etc.

Series temporales y procesos estocásticos. Características

El concepto de serie temporal se deriva de un concepto más amplio como es el de proceso estocástico. Se define un *proceso estocástico* $\{X_t\}$, para $t = 1,2,3,...$, como una colección de variables aleatorias X_t, ordenadas de acuerdo con el parámetro discreto t, que en nuestro contexto es el tiempo. Los modelos estocásticos de series temporales conciben una serie temporal dada X_t como una colección de observaciones muestrales, cada una correspondiente a una variable del proceso. Las leyes probabilísticas que rigen cualquier proceso estocástico se describen exhaustivamente y sin ambigüedades mediante las funciones de distribución de probabilidad conjunta de todos y cada uno de los vectores de variables aleatorias que se puedan formar con las variables que constituyen el proceso.

Sin embargo, para muchos fines prácticos, los procesos se suelen describir mediante sus momentos. La *media del proceso estocástico* se define por $u_t = E(X_t)$ y generalmente es una función del tiempo. La *función de autocovarianza* se define como:

$$g(t,t+k) = \mathrm{Cov}[X_t,X_{t+k}] = \mathrm{E}\{[X_t-E(X_t)][X_{t+k}-E[X_{t+k}]]\}$$
$$k = ... -3, -2, -1, 0, 1, 2, 3, ...$$

A partir de esta función se obtienen dos resultados útiles. Por una parte, para k=0 surge la *función de varianza del proceso* $g(t,t) = Var\ X_t$.

Por otra parte la *función de autocorrelación* se define como:

$$h(t,t+k) = g(t,t+k)/\ [g(t,t)g(t+k,t+k)]^{1/2}$$
$$k = ... -3, -2, -1, 0, 1, 2, 3, ...$$

Procesos estocásticos estacionarios. Funciones de autocorrelación y autocorrelación parcial

Otro concepto importante en los procesos estocásticos es el de estacionariedad. Un proceso estocástico es *estacionario en sentido estricto* si los vectores $[X_{t1}, X_{t2}, .., X_{tn}]$ y $[X_{t1+s}, X_{t2+s}, ..., X_{tn+s}]$ poseen la misma función de distribución de probabilidad, independientemente de s, para cualquier n dado. La definición de estacionariedad en sentido estricto implica que las características del proceso estocástico no sufren alteración en tiempos históricamente diferentes. Esta condición es quizá demasiado fuerte para imponer en la práctica. Un proceso es *estacionario en sentido amplio* (o estacionario de segundo orden, o de covarianza estacionaria, o débilmente estacionario) cuando se verifica que $u_t = u < \infty$ y $g(t,t+k) = g_k < \infty$, lo que significa que la media del proceso es constante (no depende del tiempo) y la autocovarianza es solo función del lapso temporal considerado, y no del tiempo histórico. Los momentos de orden superior pueden variar con el tiempo. En el caso de procesos con función de distribución de probabilidad normal, la estacionariedad en sentido amplio implica la estacionariedad en sentido estricto.

La *función de autocorrelación* FAC en procesos estacionarios es $h_k = g_k / g_0 = Cov(X_t,X_{t+k}/V(X_t)$ $k = ...-3,-2,-1,0,1,2,3...$Para procesos reales se cumple además que $g_0>0$, $g_k = g_{-k}$, $h_k = h_{-k}$, $h_0 = 1$ y $|h_k|$ menor o igual que 1. La representación gráfica con h_k en ordenadas y k en abscisas se denomina *correlograma del proceso*. La función de autocorrelación de las series estacionarias disminuye sensiblemente a medida que aumenta el desfase temporal k. Esto no suele ocurrir en las series no estacionarias.

En las aplicaciones prácticas, en las que se dispone de ciertas observaciones, X_t ($t = 1,2,\ldots,T$), relativas a un proceso estocástico que se supone estacionario, la media del proceso se estima mediante:

$$\overline{X} = \sum_{t=1}^{T} \frac{X_t}{T}$$

Análogamente, la función de autocorrelación, h_k, se estima mediante la *función de autocorrelación muestral o función de autocorrelación estimada*, que se define por:

$$r_k = \frac{\sum_{t=1}^{T}(X_t - \overline{X})(X_{t-k} - \overline{X})}{\sum_{t=1}^{T}(X_t - \overline{X})^2} \quad k = 1, 2, ...$$

La representación gráfica de r_k, denominada *correlograma muestral*, constituye un instrumento de análisis de series temporales de gran interés práctico. Para obtener correlogramas debe partirse en la práctica de muestras de tamaño suficientemente grande (al menos 50 observaciones). La función de autocorrelación muestral no se puede calcular cuando $k > T+1$, y en la práctica no debe calcularse para $k > T/4$.

Otro concepto que tiene importancia en la teoría de series temporales es el de ruido blanco. Un proceso *puramente aleatorio (ruido blanco),* se define por las condiciones:

$$u = E(X_t) = 0,\ g_0^2 = var(X_t) = \sigma,\ g_k = cov[X_t, X_{t+k}] = 0 \quad k = ..., -3,-2,-1,0,1,2,3,...$$

En este tipo de procesos, puramente aleatorios, el correlograma se reduce a un segmento de longitud unitaria sobre el eje de ordenadas.

Otro concepto muy útil en el análisis de series temporales es la *función de autocorrelación parcial* FACP de una serie temporal. El primer término de la función de autocorrelación parcial, que vamos a denotar por Φ_{11}, puede estimarse transformando la serie X_t en desviaciones respecto a su media muestral $Y_t = X_t - \overline{X}$ y a continuación estimando una regresión de Y_t sobre Y_{t-1}. La pendiente estimada de esta regresión es Φ_{11}. El modelo de regresión es $Y_t = \Phi_{11}Y_{t-1} + u_t$. Además, el primer valor de la función de autocorrelación parcial Φ_{11} es precisamente igual al primer valor de la función de autocorrelación. Esta es una propiedad de las funciones de autocorrelación de todo proceso estocástico estacionario.

El segundo valor de la función de autocorrelación parcial, Φ_{22}, se estima mediante una regresión de Y_t sobre Y_{t-1} e Y_{t-2}. El modelo de regresión es $Y_t = \Phi_{21}Y_{t1} + \Phi_{22}Y_{t-2} + u_t$. El tercer valor de la función de autocorrelación parcial, Φ_{33}, se estima mediante una regresión de Y_t sobre Y_{t-1}, Y_{t-2} e Y_{t-3}. El modelo de regresión es $Y_t = \Phi_{31}Y_{t-1} + \Phi_{32}Y_{t-2} + \Phi_{33}Y_{t-3} + u_t$.

Vemos pues que la función de autocorrelación parcial puede estimarse mediante una serie de regresiones, cada una de las cuales contiene como variable explicativa un retardo más que la anterior, y de la que nos vamos quedando en cada caso con los coeficientes estimados en los retardos más altos: Φ_{11}, Φ_{22}, Φ_{33}, ..., que son así los valores estimados de la función de autocorrelación parcial. Otra posibilidad de obtener la función de autocorrelación parcial estimada es mediante fórmulas recursivas, utilizando la función de autocorrelación previamente estimada y utilizando las ecuaciones de Yule-Walker. A veces se suele denominar *correlograma* a la representación gráfica de las funciones de autocorrelación y autocorrelación parcial.

Series temporales estacionarias. Detección de la estacionariedad

Muy pocas series temporales reales del mundo económico son *estacionarias*. La mayoría suelen presentar tendencia, suelen tener varianza no constante y también suelen presentar variaciones estacionales.

La presencia de variaciones estacionales se traduce en una variabilidad de la media del proceso, lo que es contrario a la hipótesis de estacionariedad. Pero, normalmente, es posible transformar muchas series económicas reales no estacionarias en otras aproximadamente estacionarias, sometiéndolas a operaciones algebraicas adecuadas. A las series no estacionarias que presentan una tendencia lineal se las somete a la transformación $Z_t = X_t - X_{t-1}$ para convertirlas en estacionarias (en media). Si X_t muestra una tendencia lineal, la primera diferencia de la serie, Z_t, ya no tendrá esa tendencia. En este caso se dice que X_t es una *serie temporal homogénea de primer orden o integrada de primer orden* y se denota por $I(1)$.

La eliminación de una tendencia cuadrática puede conseguirse mediante doble diferenciación. Esta operación se realiza en dos etapas, primero se obtiene $W_t = X_t - X_{t-1}$ y, si sigue existiendo tendencia, se obtiene $Z_t = W_t - W_{t-1}$. Si Z_t ya no incorpora tendencia (es estacionaria), se dice que X_t es una serie temporal homogénea de segundo orden I(2). Análogamente una tendencia de orden p puede eliminarse llevando a cabo una diferenciación de orden p dando lugar a una serie homogénea o integrada $I(p)$ de orden p.

Si hay duda sobre diferenciar o no, o sobre cuántas veces hay que diferenciar, se calcula la varianza de la serie original y de la serie sometida a diferentes diferenciaciones, tomando como diferenciación adecuada aquella para lal que la varianza es mínima. El método es tanto más adecuado cuanto mayor sea la diferencia entre las varianzas anteriores. La *sobrediferenciación* suele evitarse observando si en la parte de medias móviles alguna raíz es próxima a la unidad.

Si X_t muestra una tendencia exponencial, puede eliminarse la tendencia hallando primero el logaritmo de la serie, y luego la diferencia primera de la nueva serie así calculada. La serie $Z_t = LnX_t - LnX_{t-1}$ puede tener la tendencia eliminada.

La estacionariedad en varianza suele corregirse aplicando logaritmos o una transformación más general como la de Box-Cox. La *transformación de Box-Cox* consigue estabilizar la varianza de una serie temporal (*serie estacionaria en varianza*) y aproximar su distribución a una normal. Si X_t es la serie temporal inicial, la transformación viene dada por:

$$\begin{cases} Z_t = \dfrac{(X_t + l_2)^{l_1} - 1}{l_1 g^{l_1 - 1}} & si \quad l_1 \neq 0 \quad y \quad X_t > -l_2 \\ Z_t = gLn(X_t + l_2) & si \quad l_1 = 0 \quad y \quad l_2 < 0 \end{cases}$$

donde g es la media geométrica simple de $X_t + l_2$. El primer parámetro l_1 gobierna la fuerza de la transformación. Para l_1=1 tenemos la serie original X_t y l_2 se elige de forma que $X_t + l_2$ sea siempre positiva. Por tanto l_2 será cero si trabajamos con datos positivos e igual en valor absoluto al valor más negativo observado, en otro caso.

La transformación de Box_Cox es realmente una familia de transformaciones dependiente del parámetro l_1, que incluye como casos particulares la transformación logarítmica (l_1=0), la raíz cuadrada (l_1=1/2) y la inversa o recíproca (l_1=-1).

Una variante más sencilla de la transformación de Box- Cox es la siguiente:

$$\begin{cases} Z_t = X_t^l & si \quad l \neq 0 \quad y \quad -1 \leq l \leq 1 \\ Z_t = Ln(X_t) & si \quad l = 0 \end{cases} \quad \text{o también} \quad \begin{cases} Z_t = \dfrac{X_t^l - 1}{l} & si \quad l \neq 0 \\ Z_t = Ln(X_t) & si \quad l = 0 \end{cases}$$

Se observa que para l = -1 tenemos la transformación recíproca, para l = -1/2 tenemos la recíproca de la raíz cuadrada, para l = 0 tenemos la logarítmica, para l = 1/2 tenemos la raíz cuadrada y para l = 1 tenemos la identidad.

La *eliminación de las variaciones estacionales, para inducir la estacionariedad*, suele hacerse casi siempre, mediante la diferenciación estacional. Si los datos son mensuales, la diferenciación estacional de la serie temporal X_t, consiste en calcular $Z_t = X_t - X_{t-12}$. Con datos trimestrales calcularíamos $Z_t = X_t - X_{t-4}$. Si después de efectuar esta transformación la serie sigue presentando evidencias de variaciones estacionales, es posible aplicar de nuevo el procedimiento, es decir, calcular las diferencias de segundo orden, y así sucesivamente.

Para detectar rápidamente la estacionariedad se puede utilizar directamente el *gráfico de la serie*. Se divide el campo de variación total de la serie en varios intervalos calculándose para cada uno de ellos la media y la varianza. Si existe estacionalidad se toma como longitud del intervalo la del periodo estacional. Para ver si la serie es *estacionaria en media* basta comprobar que las medias de los intervalos no fluctúen mucho. Para ver si la serie es *estacionaria en varianza* basta comprobar que las varianzas de los intervalos son estables (no cambian bruscamente) y se mantienen en una franja estrecha. La Figura 6-4 ilustra estos conceptos. Otro criterio para detectar la estacionariedad en varianza es el gráfico *rango-media de Box-Cox* (Figura 6-5), consistente en representar los puntos (media, rango) para todos los intervalos en que se ha dividido la serie. Si los puntos del gráfico son ajustables a una recta con pendiente positiva no hay estacionariedad en varianza (será necesario tomar logaritmos en la serie original si $\lambda = 0$ y elevar la serie a un exponente fraccionario para otro valor de λ distinto de uno). Si el gráfico no tiene tendencia definida o es ajustable a una recta paralela al eje de abscisas hay estacionariedad en varianza ($\lambda = 1$).

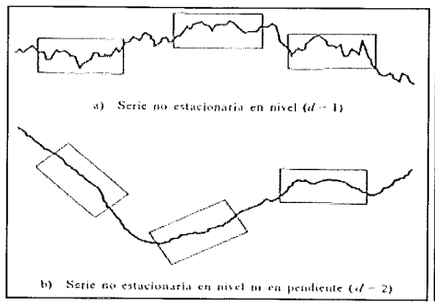


Figura 6-4

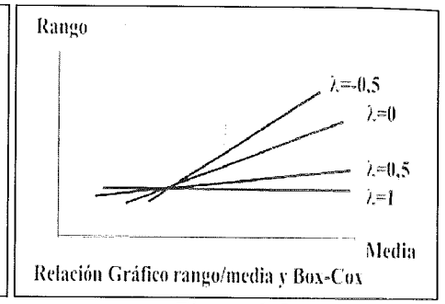


Figura 6-5

Otro criterio para detectar la estacionariead es el *criterio de la función de autocorrelación estimada*. Si los coeficientes de la FAC no decaen rápidamente hay un indicio claro de falta de estacionariedad en media, lo que nos llevaría a tomar primeras diferencias en la serie original.

Un criterio formal para detectar la estacionariedad son los *contrastes de raíces unitarias (ADF, Phillips Perron,* etc*)*, que se estudiarán en capítulos posteriores.

Un proceso *puramente aleatorio (ruido blanco),* se define por las condiciones $u = E(X_t) = 0$, $g_0^2 = var(X_t) = \sigma$; $g_k = cov[X_t, X_{t+k}] = 0 \quad k = ..., -3,-2,-1,0,1,2,3,...$ En este tipo de procesos, puramente aleatorios, el correlograma se reduce a un segmento de longitud unitaria sobre el eje de ordenadas.

MODELOS AUTORREGRESIVOS AR(p)

Un modelo autorregresivo (AR) describe una clase particular de proceso en el que las observaciones en un momento dado son predecibles a partir de las observaciones previas del proceso más un término de error. El caso más simple es el ARIMA(1,0,0), o AR(1) o de primer orden, cuya expresión matemática es:

$$X_t = \Phi_1 X_{t-1} + a_t$$

El proceso autorregresivo de orden p, representado por ARIMA(p,0,0), o simplemente por AR(p) toma la forma:

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + ... + \Phi_p X_{t-p} + a_t$$

que puede ponerse, mediante el operador de cambio retroactivo B, en la forma:

$$(1 - \Phi_1 B - \Phi_2 B^2 - ... - \Phi_p B^p)\, Xt = a_t \qquad B^k(X_t) = X_{t-k}$$

Un proceso autorregresivo AR(p) es *estacionario* si las raíces del polinomio en B dado por: $1 - \Phi_1 B - \Phi_2 B^2 - \ldots - \Phi_p B^p$ caen fuera del círculo unidad. Esa condición es equivalente a que las raíces de la ecuación: $x^p - \Phi_1 x^{p-1} - \Phi_2 x^{p-2} - \ldots - \Phi_{p-1} x - \Phi_p = 0$ sean todas inferiores a uno en módulo. Un proceso autorregresivo siempre es invertible.

La varianza de un proceso AR(1) es: $g_0 = \dfrac{\sigma_a^2}{1-\Phi_1}$

La función de autocovarianza de un proceso AR(1) es:

$$g_k = \Phi_1^k \frac{\sigma_a^2}{1-\Phi_1} \quad k \geq 1$$

La función de autocorrelación de un proceso AR(1) es :

$$h_k = \Phi_1^k \quad k \geq 1$$

La función de autocorrelación parcial de un proceso AR(1) es:

$$h_{kk} = \begin{cases} \Phi_1 & para \quad j = 1 \\ 0 & para \quad j > 1 \end{cases}$$

La varianza de un proceso AR(2) es: $g_0 = \Phi_1 g_1 + \Phi_2 g_2 + \sigma_a^2$

La función de autocovarianza de un proceso AR(2) es:

$$g_k = \Phi_1 g_{k-1} + \Phi_2 g_{k-2} \quad k \geq 1$$

La función de autocorrelación de un proceso AR(2) es :

$$h_k = \Phi_1 h_{k-1} + \Phi_2 h_{k-2} \quad k \geq 1$$

La función de autocorrelación parcial de un proceso AR(2) es:

$$h_{kk} = \begin{cases} h_1 = \dfrac{\Phi_1}{1-\Phi_2} & para \quad j = 1 \\ \dfrac{h_2 - h_1^2}{1-h_1^2} = \Phi_2 & para \quad j = 2 \\ 0 & para \quad j > 2 \end{cases}$$

La varianza de un proceso AR(p) es: $g_0 = \Phi_1 g_1 + \Phi_2 g_2 + \cdots \Phi_p g_p + \sigma_a^2$

La función de autocovarianza de un proceso AR(p) es:

$$g_k = \Phi_1 g_{k-1} + \Phi_2 g_{k-2} + \cdots + \Phi_p g_{k-p} \quad k \geq 1$$

La función de autocorrelación de un proceso AR(p) es:

$$h_k = \Phi_1 h_{k-1} + \Phi_2 h_{k-2} + \cdots + \Phi_p h_{k-p} \quad k \geq 1$$

La función de autocorrelación parcial de un proceso AR(p) es:

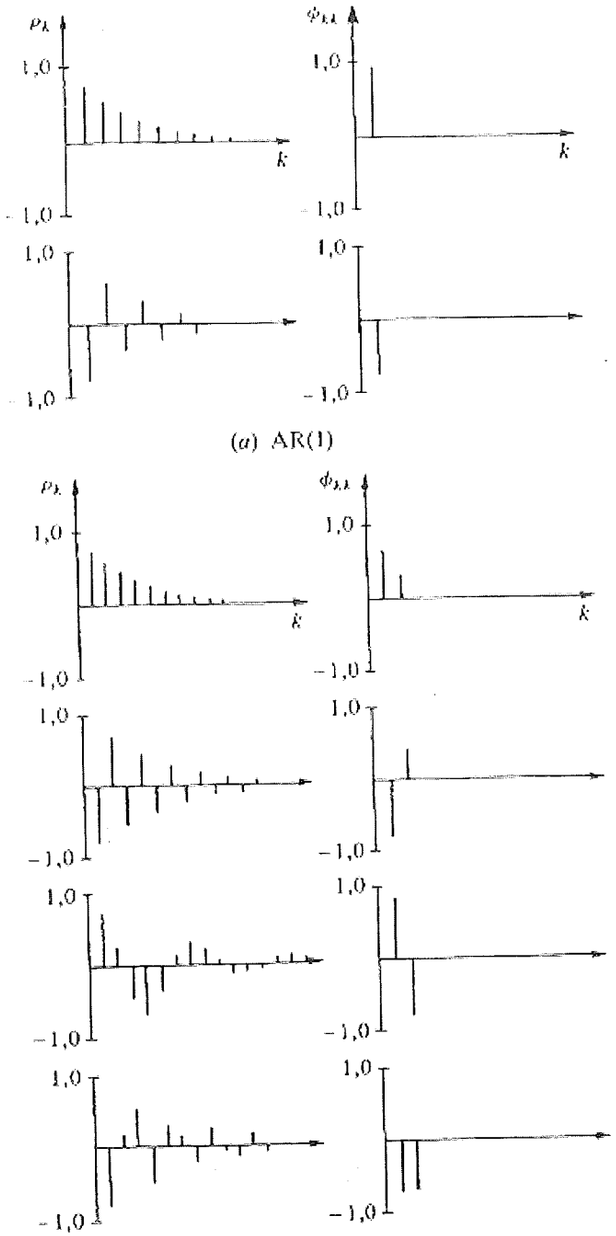
$$h_{kk} = \begin{cases} h_1 & para \quad j = 1 \\ \dfrac{h_2 - h_1^2}{1-h_1^2} & para \quad j = 2 \\ \dfrac{\begin{vmatrix} 1 & h_1 & \cdots & h_{p-2} & h_1 \\ h_1 & 1 & \cdots & h_{p-3} & h_2 \\ \vdots & \vdots & & \vdots & \vdots \\ h_{p-1} & h_{p-2} & \cdots & h_1 & h_p \end{vmatrix}}{\begin{vmatrix} 1 & h_1 & \cdots & h_{p-2} & h_{p-1} \\ h_1 & 1 & \cdots & h_{p-3} & h_{p-2} \\ \vdots & \vdots & & \vdots & \vdots \\ h_{p-1} & h_{p-2} & \cdots & h_1 & 1 \end{vmatrix}} & para \quad j = p \\ 0 & para \quad j > p \end{cases}$$

En la Figura 2-3 se observan las funciones de autocorrelación (izquierda) y autocorrelación parcial (derecha) para procesos AR(1) y AR(2).

MODELOS DE MEDIAS MÓVILES MA(q)

Un modelo de *medias móviles (MA)* también describe una serie temporal estacionaria. En este modelo el valor actual puede predecirse a partir de la componente aleatoria de este momento y, en menor medida, de los impulsos aleatorios anteriores. El modelo ARIMA(0,0,1), también denotado por MA(1), viene dado por la expresión:

$$X_t = a_t - v_1 a_{t-1}$$



(a) AR(1)

(b) AR(2)

Figura 6-6

El proceso de medias móviles de orden q, representado por ARIMA(0,0,q), o también por MA(q), viene dado por la expresión:

$$X_t = a_t - \nu_1 a_{t-1} - \nu_2 a_{t-2} - .... - \nu_q a_{t-q}$$

que puede ponerse, mediante el operador de cambio retroactivo B, en la forma:

$$X_t = (1 - \nu_1 B - \nu_2 B^2 - .... - \nu_q B^q)\, a_t$$

Un proceso de medias móviles es siempre estacionario.

Un proceso de medias móviles $MA(q)$ es *invertible* si las raíces del polinomio en B definido por: $1 - \nu_1 B - \nu_2 B^2 - .... - \nu_q B^q$ caen fuera del círculo unidad. Esta condición es equivalente a que las raíces de la ecuación $x^q - \Phi_1 x^{q-1} - \Phi_2 x^{q-2} - ... - \Phi_{q-1} x - \Phi_q = 0$ sean todas inferiores a uno en módulo.

La varianza de un proceso MA(1) es: $g_0 = \sigma_a^2 (1 + \nu_1^2)$

La función de autocovarianza de un proceso M(1) es:

$$g_k = \begin{cases} -\nu_1 \sigma_a^2 & para \quad k = 1 \\ 0 & para \quad k > 1 \end{cases}$$

La función de autocorrelación de un proceso MA(1) es :

$$h_k = \begin{cases} \dfrac{-\nu_1}{1 + \nu_1^2} & para \quad k = 1 \\ 0 & para \quad k > 1 \end{cases}$$

La función de autocorrelación parcial de un proceso MA(1) es:

$$h_{kk} = \frac{-\nu_1^k (1 - \nu_1^2)}{1 - \nu_1^{2(k+1)}} \quad para \quad k \geq 1$$

La varianza de un proceso MA(2) es: $g_0 = \sigma_a^2 (1 + \nu_1^2 + \nu_2^2)$

La función de autocovarianza de un proceso MA(2) es:

$$g_k = \begin{cases} -(\nu_1 + \nu_1\nu_2)\sigma_a^2 & para \quad k=1 \\ -\nu_2\sigma_a^2 & para \quad k=2 \\ 0 & para \quad k>2 \end{cases}$$

La función de autocorrelación de un proceso MA(2) es :

$$h_k = \begin{cases} \dfrac{-\nu_1 + \nu_1\nu_2}{1+\nu_1^2+\nu_2^2} & para \quad k=1 \\ \dfrac{-\nu_2}{1+\nu_1^2+\nu_2^2} & para \quad k=2 \\ 0 & para \quad k>2 \end{cases}$$

La función de autocorrelación parcial de un proceso MA(2) es:

$$h_{kk} = \begin{cases} h_1 & para \quad k=1 \\ \dfrac{h_2 - h_1^2}{1-h_1^2} & para \quad k=2 \\ \dfrac{h_1^3 - h_1h_2(2-h_2)}{1-h_2^2-2h_1^2(1-h_2)} & para \quad k=3 \\ \cdots \end{cases}$$

La varianza de un proceso MA(q) es: $g_0 = \sigma_a^2(1+\nu_1^2+\nu_2^2+\cdots+\nu_q^2)$

La función de autocovarianza de un proceso MA(q) es:

$$g_k = \begin{cases} (-\nu_k + \nu_1\nu_{k+1} + \cdots + \nu_{q-k}\nu_q)\sigma_a^2 & para \quad k=1,2,\cdots,q \\ 0 & para \quad k>q \end{cases}$$

La función de autocorrelación de un proceso MA(q) es:

$$h_k = \begin{cases} \dfrac{-\nu_k + \nu_1\nu_{k+1} + \cdots + \nu_{q-k}\nu_q}{1+\nu_1^2+\cdots+\nu_q^2} & para \quad k=1,2,\cdots,q \\ 0 & para \quad k>q \end{cases}$$

La función de autocorrelación parcial de un proceso MA(q) es:

$$h_{kk} = \begin{cases} h_1 & para \quad k=1 \\ \dfrac{h_2 - h_1^2}{1-h_1^2} & para \quad k=2 \\ \dfrac{h_1^3 - h_1h_2(2-h_2) + h_3(1-h_1^2)}{1-h_2^2-2h_1^2(1-h_2)} & para \quad k=3 \\ \cdots \end{cases}$$

En la Figura 6-7 se observan las funciones de autocorrelación (izquierda) y autocorrelación parcial (derecha) para procesos MA(1) y MA(2).

MODELOS ARMA(p,q)

Una extensión natural de los *modelos AR(p)* y *MA(q)* es un tipo de modelos que incluyen tanto términos autorregresivos como de medias móviles y se definen como ARMA(p,q) o también como ARIMA(p,0,q). Se representan por la ecuación:

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \ldots + \Phi_p X_{t-p} + a_t - \nu_1 a_{t-1} - \nu_2 a_{t-2} - \ldots - \nu_q a_{t-q}$$

que puede ponerse de la forma:

$$X_t - \Phi_1 X_{t-1} - \Phi_2 X_{t-2} - \ldots - \Phi_p X_{t-p} = a_t - \nu_1 a_{t-1} - \nu_2 a_{t-2} - \ldots - \nu_q a_{t-q}$$

o sea:

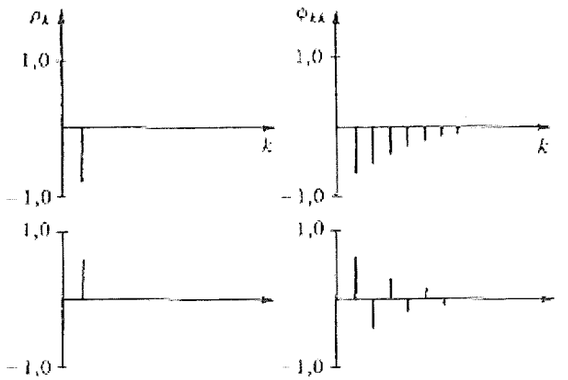
$$(1 - \Phi_1 B - \Phi_2 B^2 - \ldots - \Phi_p B^p)\, X_t = (1 - \nu_1 B - \nu_2 B^2 - \ldots - \nu_q B^q)\, a_t$$

El proceso ARMA(p,q) es estacionario si lo es su componente autorregresiva, y es invertible si lo es su componente de medias móviles. Por tanto podemos decir que un modelo ARMA(p,q) es invertible si las raíces del polinomio en B definido mediante $1 - \nu_1 B - \nu_2 B^2 - \ldots - \nu_q B^q$ caen fuera del círculo unidad. Esta condición es equivalente a que las raíces de la ecuación $x^q - \Phi_1 x^{q-1} - \Phi_2 x^{q-2} - \ldots - \Phi_{q-1} x - \Phi_q = 0$ sean todas inferiores a uno en módulo.

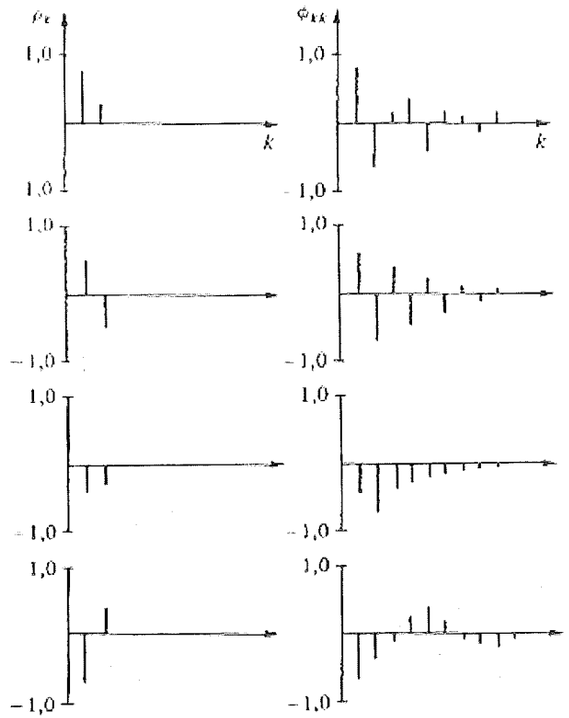
Un modelo ARMA(p,q) es estacionario si las raíces del polinomio definido por $1 - \Phi_1 B - \Phi_2 B^2 - \ldots - \Phi_p B^p$ caen fuera del círculo unidad. Esa condición es equivalente a que las raíces de la ecuación: $x^p - \Phi_1 x^{p-1} - \Phi_2 x^{p-2} - \ldots - \Phi_{p-1} x - \Phi_p = 0$ sean todas inferiores a uno en módulo.

La varianza de un proceso ARMA(1,1) es:

$$g_0 = \frac{\sigma_a^2(1+\nu_1^2 - 2\Phi_1\nu_1)}{1-\Phi_1^2}$$



(a) MA(1)



(b) MA(2)

Figura 6-7

La función de autocovarianza de un proceso ARMA(1,1) es:

$$g_k = \begin{cases} \dfrac{\sigma_a^2(1-\Phi_1\nu_1)(\Phi_1-\nu_1)}{1-\Phi_1^2} & para \quad k=1 \\ \Phi_1 g_{k-1} & para \quad k>1 \end{cases}$$

La función de autocorrelación de un proceso ARMA(1,1) es:

$$h_k = \begin{cases} \dfrac{(1-\Phi_1\nu_1)(\Phi_1-\nu_1)}{1-\Phi_1^2-2\Phi_1\nu_1} & para \quad k=1 \\ \Phi_1 h_{k-1} & para \quad k>1 \end{cases}$$

La función de autocorrelación parcial de un proceso ARMA(p,q) es:

$$h_{kk} = \begin{cases} h_1 & para \quad k=1 \\ \dfrac{h_2-h_1^2}{1-h_1^2} & para \quad k=2 \\ \dfrac{h_1^3-h_1h_2(2-h_2)+h_3(1-h_1^2)}{1-h_2^2-2h_1^2(1-h_2)} & para \quad k=3 \\ \ldots \end{cases}$$

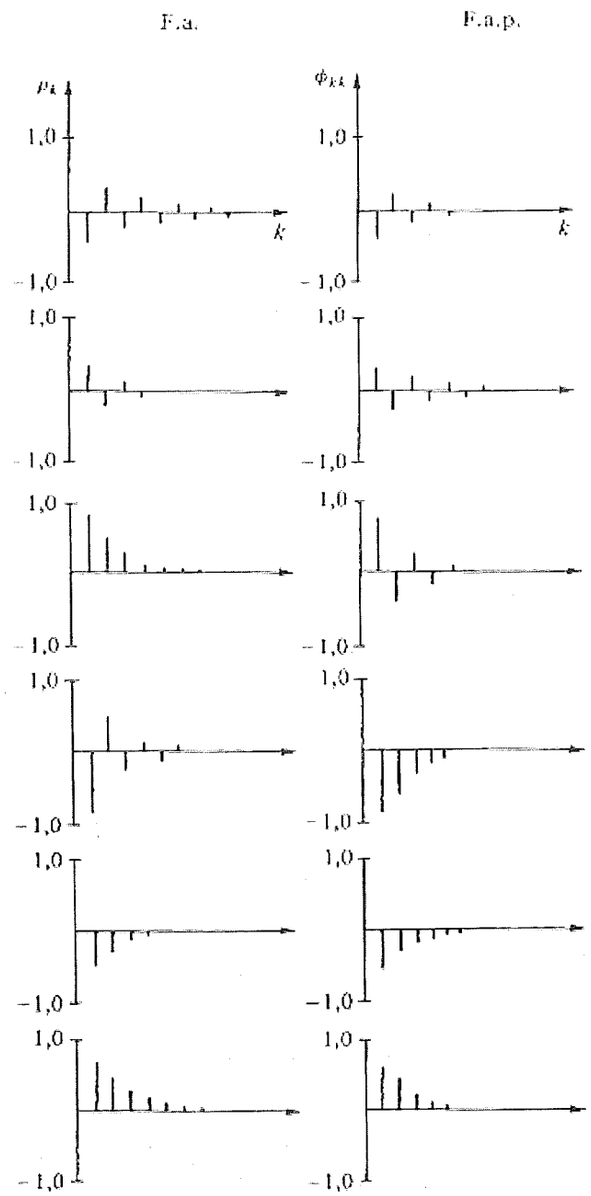
En la Figura 6-8 se observan las funciones de autocorrelación (izquierda) y autocorrelación parcial (derecha) para procesos ARMA(1,1).

MODELOS ARIMA(p,d,q)

Un modelo ARIMA(0,*d*,0) es una serie temporal que se convierte en un ruido blanco (proceso puramente aleatorio) después de ser diferenciada *d* veces. El modelo ARIMA(0,*d*,0) se expresa mediante: $(1 - B)d\ X_t = a_t$. El modelo general ARIMA(*p,d,q*) denominado proceso autorregresivo integrado de medias móviles de orden *p, d, q*, toma la siguiente expresión:

$$(1- \Phi_1 B - \Phi_2 B^2 -\ldots- \Phi_p B^p)(1-B)^d\ Y_t = (1 - \nu_1 B - \nu_2 B^2 - \ldots\ \nu_q B^q)a_t$$

Un modelo ARIMA(*p,d,q*) permite describir una serie de observaciones después de que hayan sido diferenciadas *d* veces, a fin de extraer las posibles fuentes de no estacionariedad. Esta fórmula general se puede aplicar a cualquier modelo. Si hay alguna componente *p,d,q* igual a cero, se elimina el término correspondiente de la fórmula general.



ARMA(1,1)

Figura 6-8

LA METODOLOGÍA BOX JENKINS EN MODELOS ARIMA

Box y Jenkins en su desarrollo de modelos estadísticos para series temporales fijaron distintas fases para su modelado. Básicamente estas fases se resumen en la identificación del modelo ARIMA adecuado a los datos de la serie (recogida de datos de la serie, representación gráfica, análisis de la estacionariedad, transformaciones previas adecuadas para conseguir la estacionariedad, eliminación de la tendencia si es necesario e identificación efectiva del modelo asociándolo a la estructura ARIMA adecuada), estimación del modelo previamente identificado (cálculo de los estimadores del modelo y residuales), validación del modelo (contrastes para ver si el modelo es adecuado) y predicción (selección de los periodos de predicción y cálculo de estadísticos para evaluar la capacidad predictiva).

La metodología para modelos ARIMA contempla las siguientes fases:

1. ***Recogida de datos de la serie.*** Es conveniente disponer de cincuenta o más datos, y en el caso de series mensuales, es habitual trabajar con entre seis y diez años completos de información. El mismo criterio se sigue para series con diferentes periodos estacionales.

2. ***Representación gráfica de la serie.*** Como primera tarea del proceso de identificación, para decidir sobre la estacionariedad de la serie es de gran utilidad disponer de un gráfico de la misma. A veces suelen utilizarse medias y desviaciones típicas por subperiodo para juzgar sobre la estacionariedad de la serie. Por ello es necesario calcular todo tipo de estadísticos relativos a la serie y necesarios en el proceso de identificación.

3. ***Transformación previa de la serie.*** También dentro del proceso de identificación, la transformación logarítmica es necesaria en caso de serie no estacionaria en varianza. Sin embargo, es una transformación muy frecuente, incluso en series con dispersión relativamente constante en el tiempo. Una posibilidad práctica es ensayar siempre con la serie original y en logaritmos y comprobar resultados. Puede ser necesario también utilizar cualquier tipo de transformación Box Cox para estacionarizar la serie, en cuyo caso se identificarán los parámetros adecuados para la transformación

4. ***Eliminación de la tendencia.*** La observación del gráfico de la serie nos indicará la existencia o no de tendencia. Una tendencia lineal será corregida tomando primeras diferencias, que será el caso más frecuente (d=1). Una tendencia no lineal suele llevar en la práctica al uso de dos diferencias como mucho (d=2). Estacionarizada la serie, habremos identificado el parámetro d.

5. *Identificación efectiva del modelo.* Consiste en determinar el tipo de modelo más adecuado para la serie objeto de estudio, es decir, el orden de los procesos autorregresivos p y de medias móviles q de las componentes regular y estacional. Técnicamente esta decisión se tomará en base a las funciones de autocorrelación y autocorrelación parcial. Habitualmente se terminará eligiendo entre los procesos más simples AR(1), AR(2), MA(1), MA(2) y ARMA(1,1), tanto en la parte regular como en la estacional. En caso de duda pueden seleccionarse varios modelos alternativos que serán estimados y contrastados posteriormente, para definir el modelo definitivamente adoptado.

6. *Estimación de los coeficientes del modelo.* Decidido el modelo, se procede a la estimación de sus parámetros. Dado que se trata de un procedimiento iterativo de cálculo, pueden sugerirse valores iniciales.

7. *Contraste de validez del modelo o validación.* Utilizaremos diversos procedimientos para valorar el modelo o modelos inicialmente seleccionados: contraste de significación de parámetros, covarianzas entre estimadores, coeficiente de correlación, suma de cuadrados de errores, etc.

8. *Análisis detallado de los errores.* Las diferencias históricas entre valores reales y estimados por el modelo constituyen una fuente de especial interés para una valoración final del modelo. Deberá comprobarse un comportamiento no sistemático de los mismos, así como analizarse la posible existencia de errores especialmente significativos.

9. *Selección del modelo.* En base a los resultados de las etapas anteriores, debe estarse en condiciones de decidir sobre el modelo adoptado.

10. *Predicción.* El modelo seleccionado servirá como fórmula inicial de predicción.

Identificación de modelos ARIMA

Identificar un modelo significa utilizar los datos recogidos, y cualquier información de cómo se genera la serie temporal objeto de estudio, para sugerir un conjunto reducido de posibles modelos, que tengan muchas posibilidades de ajustarse a los datos. Ante una serie temporal empírica, el investigador debe encontrar los valores p, d, q más apropiados.

Si la serie temporal presenta una tendencia, lo primero que se deber hacer es convertirla en estacionaria mediante una diferenciación de orden d. Una vez diferenciada la serie, una buena estrategia consiste en comparar los correlogramas de la función de autocorrelación (FAC) y la función de autocorrelación parcial (FACP). Esto suele ofrecer una orientación para la formulación del modelo tentativo.

Los procesos autorregresivos presentan función de autocorrelación parcial con un número finito de valores distinto de cero.

Un proceso AR(p) tiene los primeros p términos de la función de autocorrelación parcial distintos de cero y los demás son nulos (Figura 6-6). Esta afirmación es muy fuerte, y en la práctica se considera que una muestra dada proviene de un proceso autorregresivo de orden p si los términos de la función de autocorrelación parcial son casi cero a partir del que ocupa el lugar p. Un valor se considera casi cero cuando su módulo es inferior a $2/\sqrt{T}$. Los programas de ordenador construyen la franja $(-2/\sqrt{T}, 2/\sqrt{T})$ y detectan los valores de la FACP que caen fuera de ella.

Los procesos de medias móviles presentan función de autocorrelación con un número finito de valores distinto de cero. *Un proceso MA(q) tiene los primeros q términos de la función de autocorrelación distintos de cero y los demás son nulos* (Figura 6-7). Estas propiedades son muy importantes con vistas a la identificación de un proceso mediante el análisis de las funciones de autocorrelación y autocorrelación parcial.

Para modelos ARMA(p,q), los primeros valores de la función de autocorrelación no tienen patrón fijo y van seguidos de una mezcla de oscilaciones sinusoidales o exponenciales amortiguadas. Asimismo, los primeros valores de la función de autocorrelación parcial no tienen patrón fijo, aunque suelen decrecer, y van seguidos de una mezcla de oscilaciones sinusoidales y exponenciales amortiguadas. La Figura 6-8 muestra estos patrones para distintos procesos ARIMA(1,1).

Podemos resumir los pasos para la identificación de un modelo de series temporales de la siguiente forma:

1. Decidir si X_t necesita ser transformada para eliminar la no estacionariedad en media o la no estacionariedad en varianza (heteroscedasticidad). Puede ser conveniente usar logaritmos de la serie o aplicar la transformación de Box-Cox.

2. Determinación del grado de diferenciación adecuado d. En general *la falta de estacionariedad, se manifiesta en que los coeficientes de la función de autocorrelación estimada tienden a decrecer muy lentamente.* La cuestión es, sin embargo, ¿cuán lentamente ha de ser el decrecimiento de los coeficientes de la función de autocorrelación parcial para que el proceso sea estacionario? En general, sólo ocasionalmente los datos económicos del correlograma dejarán de decrecer tras las primeras diferencias, y en este caso serían necesarias segundas diferencias. Una diferenciación superflua sólo sirve para alterar el esquema de autocorrelación evidente en una serie estacionaria y complicarlo innecesariamente.

3. Decidir los valores de p y q, y si existe una componente estacional, decidir los órdenes de los operadores estacionales P y Q. Para este apartado se utilizan las funciones de autocorrelación y autocorrelación parcial según el siguiente cuadro:

Proceso	Función de autocorrelación	Función de autocorrelación parcial
MA(q)	Sólo los q primeros coeficientes son significativos. El resto se anulan bruscamente (coef. 0 para retardo>q)	Decrecimiento rápido exponencial atenuado u ondas sinusoidales
AR(p)	Decrecimiento rápido exponencial atenuado u ondas sinusoidales	Sólo los p primeros coeficientes son significativos. El resto se anulan bruscamente (coef. 0 para retardo>p)
ARMA	Los coeficientes no se anulan bruscamente	Los coeficientes no se anulan bruscamente
ARIMA(p,d,q)	Comportamiento irregular en los retardos (1,...,q) con q picos. Decrecimiento para retardos posteriores a q	Decrece (aproximadamente con exponenciales atenuados y ondas sinusoidales). No cero pronto

Estimación de modelos ARIMA(p,d,q)

El criterio que suele utilizarse es obtener los parámetros de manera que la suma cuadrática de los errores sea lo menor posible. Si representamos el proceso ARIMA(p,d,q) de la forma $\Phi(B)\ X_t = \nu(B)\ a_t$ los errores del modelo pueden expresarse de la forma $a_t = \Phi^{-1}(B)\ \Phi(B)\ a_t$ de forma que el objetivo es encontrar el vector de parámetros $\Phi = \Phi\,(\Phi_1,\ldots,\ \Phi_p)$ y $\nu = (\nu_1,\ldots,\nu_p)$ que minimice la suma de cuadrados de los errores $\sum_t a_t^2 = S(\Phi,\nu)$.

La estimación es complicada ya que la ecuación es no lineal en los parámetros. Debemos, pues, utilizar un método iterativo de estimación no lineal, como por ejemplo el de Marquardt. Para comenzar el algoritmo necesitamos estimaciones preliminares de los parámetros, que se obtienen mediante el método de los momentos.

Diagnóstico, validación o contraste de modelos ARIMA(p,d,q)

Box y ***Jenkins*** sugirieron un número considerable de tests para verificar si el modelo elegido se ajusta correctamente al conjunto de datos dado. Uno de ellos, conocido como sobreparametrización, consiste en ajustar un modelo de orden superior al elegido y comprobar si los parámetros son significativamente distintos de cero.

Por otra parte, si el modelo aproxima satisfactoriamente a la serie observada, los residuos deben tender a comportarse como ruido blanco, lo cual se comprobaría mediante las *funciones de autocorrelación de los residuos* (*FAC y FACP*). Dichas funciones de autocorrelación deben ser nulas en todo su recorrido, excepto en cero.

Si el modelo no aproxima satisfactoriamente a la serie observada, los residuos se comportarán como un ruido autocorrelado, problema análogo al encontrado en los modelos econométricos con perturbaciones autocorrelacionadas. Por ello, deben emplearse contrastes como el de Durbin-Watson (para la autocorrelación de primer orden) o el de Wallis (para la de cuarto orden).

Otros tests, aplicados a los residuos, van encaminados a comprobar si los residuos obtenidos son consistentes con el supuesto de ruido blanco (aleatorios).

Box y ***Pierce*** proponen el estadístico $Q = \sum_{k=1}^{m} r_k^2$ donde r_k viene definido por:

$$r_k = \sum_{t=k+1}^{n} a_t a_{t-k} \Big/ \sum_{t=1}^{n} a_t^2 \quad a_t = \text{residuos estimados, } n = \text{número de observaciones.}$$

Bajo el supuesto de que m es suficientemente grande, *Box* y *Pierce* demuestran que el estadístico Q se distribuye como una *Chi-cuadrado* con m-p-q grados de libertad. La hipótesis de que los residuos son un ruido blanco se rechaza en general para valores de Q muy altos. Más concretamente, se halla la región crítica a nivel α, calculando un valor I que cumpla $P(Q>I)= \alpha$. Si el valor del estadístico Q cae dentro de la región crítica, que es $\{Q>I\}$, entonces se rechaza la hipótesis nula de que los residuos son un ruido blanco. Si cae fuera se acepta la hipótesis nula. El valor de m es arbitrario, pero conviene tomarlo lo más elevado posible.

Para valores de m no muy grandes, ***Ljung*** y ***Box*** proponen un estadístico alternativo:

$$Q' = n(n+2)\sum_{k=1}^{m} r_k^2 \big/ (n-k)$$

que también se distribuye como una *Chi-cuadrado* con m-p-q grados de libertad. Se halla la región crítica a nivel α, calculando un valor I que cumpla $P(Q'>I)= \alpha$. Si el valor del estadístico Q' cae dentro de la región crítica, que es $\{Q'>I\}$, entonces se rechaza la hipótesis nula de que los residuos son un ruido blanco. Si cae fuera se acepta la hipótesis nula.

Un diagnóstico completo también surge de la ***inspección del gráfico de los residuos***. Si los residuos provienen de un proceso de ruido blanco, deben ser incorrelacionados entre sí, lo que les hará alternar en signo, sin ningún criterio obvio. Por el contrario, rachas de residuos consecutivos de un mismo signo son, en general, un indicativo de mala especificación del modelo, bien por ser una indicación de autocorrelación de los residuos o por indicar no estacionariedad de los mismos. Si los residuos representados contra el índice tiempo t, es decir si el grafo (t,a_t), tiene una tendencia conocida, puede haber heteroscedasticidad de los residuos. Aquí se pueden aplicar todos los contrastes de aleatoriedad, autocorrelación, heteroscedasticidad, falta de linealidad y no normalidad de los residuos.

El ***periodograma de los residuos*** debe presentar amplitudes destacables en casi toda la gama de frecuencias. El ***periodograma acumulativo de los residuos*** debe producir una curva de amplitudes sobre la recta de reposo sin presentar patrones de oscilación en ninguna zona de frecuencias.

También existen métodos de otro tipo para contrastar la bondad del modelo univariante estimado. Conviene estimar el modelo excluyendo algunas observaciones al final de la muestra. Si esto provoca una variación sensible en los valores estimados de los parámetros podría indicar una variación reciente de la estructura estocástica subyacente, lo que desaconsejaría el modelo para fines predictivos.

Por otro lado, los modelos ARMA(p,q) deben cumplir las condiciones de estacionariedad e invertibilidad. Por tanto, si representamos el proceso ARMA(p,q) de la forma $\Phi(B)\ X_t = v(B)a_t$ y alguna de las raíces de las ecuaciones $\Phi(B) = 0$ y $v(B)=0$ es menor que uno en módulo, el modelo es rechazable.

Si alguna de las raíces de la ecuación $\Phi(B) = 0$ es muy próxima a la unidad, la serie original puede estar subdiferenciada y precisará alguna diferenciación adicional. Si alguna de las raíces de la ecuación $v(B) = 0$ es muy próxima a la unidad, la serie original puede estar sobrediferenciada. Si coincide una raíz de ambas ecuaciones, se puede cancelar un orden en el proceso, pasando a un ARMA(p-1,q-1).

Predicción en modelos ARIMA

Los modelos ARIMA proporcionan no solamente una predicción puntual, sino la distribución de probabilidad completa para los valores futuros de la serie. Considerando una predicción óptima a aquélla con un error cuadrático medio de predicción mínimo, trataríamos de elegir nuestra predicción a horizonte l, $Z_t(l)$, tal que $E[e_t^2(l)] = E\{\ [X_{t+1}-Z_t(l)]^2\ \}$ fuese mínimo. En general se puede demostrar que dicha predicción viene dada por la esperanza condicionada de X_{t+l}, es decir:

$$Z_t(l) = E[X_{t+l}/X_t,X_{t-1},\ldots,X_1]$$

El cálculo real de la predicción $Z_t(l)$ puede hacerse de forma recursiva utilizando el modelo ARIMA estimado, de forma que si escribimos el modelo como

$$d_t = \Phi_1\, d_{t-1} +\ldots+ \Phi_p d_{t-p} + a_t - v_1 a_{t-1} - \ldots - v_q\, a_{t-q}$$

donde d_t es la diferencia de orden d de X_t (supuesto X_t no estacionaria y convertible en estacionaria mediante un proceso de d diferenciaciones consecutivas).

Para calcular la predicción $Z_t(l)$, se comienza calculando la estimación de $d_t(1)$ como la esperanza condicionada de d_{t+1}, y posteriormente se calcula la estimación de $d_t(2)$, y así sucesivamente hasta calcular la estimación de $d_t(l)$. Una vez que la serie d_t ha sido predicha, podemos obtener una predicción de X_t sumando d_t d veces. Para calcular la predicción $Z_t(l)$ utilizamos la siguiente fórmula:

$$Z_t(l)= \Phi_l d_t + \Phi_{l+1}\, d_{t-1} + \Phi_{l+2}\, d_{t-2} +\ldots= Z_{t+l}$$

SERIES TEMPORALES ESTACIONALES. DETECCIÓN DE LA ESTACIONALIDAD

Las series estacionales presentan oscilaciones que se producen con un periodo igual o inferior a un año, y que se reproducen de manera reconocible en los diferentes años. El motivo principal que induce a estudiar la componente estacional es que en la inmensa mayoría de las series económicas dicha componente provoca una distorsión de su verdadero movimiento. Para eliminar estas distorsiones y captar el movimiento real de la serie, es necesario eliminar las oscilaciones estacionales desestacionalizando la serie.

Cuando se representa una serie temporal mediante y_t, se suponen todas las observaciones ordenadas una detrás de otra tal y como se van produciendo (t = 1, 2, ..., T). Cuando representamos una serie temporal por y_{ik} , estamos considerando explícitamente el año i (i = 1, 2, ..., N) y la estación del año k (k = 1, 2, ..., m). Cuando la estación es el año m = 12 , y cuando es el trimestre, m = 4. Siempre se tiene que $T = Nm$.

Para detectar la estacionalidad pueden utilizarse los siguientes caminos:

- El *gráfico de la serie* da una idea de los posibles periodos estacionales.
- El *gráfico de las subseries estacionales* identifica gráficamente los periodos estacionales presentando secciones sucesivas de los mismos.
- El *gráfico de las subseries anuales* valida gráficamente los periodos estacionales presentando comportamientos paralelos para cada estación.
- Las *funciones de autocorrelación y autocorrelación parcial estimadas* también validan los periodos estacionales de acuerdo a las siguientes consideraciones:
  - ➢ Los coeficientes de la FAC para retardos múltiplos del periodo estacional de la serie deben ser significativamente distintos de cero
  - ➢ Para una cantidad grande de retardos la FAC se configura en forma de abanico que completa su ciclo girando sobre el eje de abscisas para una cantidad de retardos igual al periodo estacional. La FACP debe presentar estructura de coeficientes significativos para retardos periódicos (largos)
  - ➢ La FAC y la FACP deben considerase a la vez, pues a veces intercambian sus papeles en el comportamiento estacional.
- El ***periodograma***, que es una figura que transforma la serie temporal de su dominio natural (que es el tiempo) al dominio de las frecuencias (a los valores de la serie se les aplican transformaciones de Fourier). Se representan frecuencias en el eje X y amplitudes en el eje Y. Respecto del periodograma tendremos en cuenta lo siguiente:

- Si no hay picos destacables en el periodograma no hay estacionalidad.
- Cada pico destacable identifica un periodo que incluso puede ser un ciclo.
- A cada amplitud destacable le corresponde una frecuencia cuya inversa es el periodo estacional o el ciclo, con lo que *el periodograma identifica la longitud del periodo estacional y en su caso el ciclo.*
- Las amplitudes más fuertes, correspondientes a valores más bajos de las frecuencias suelen corresponder a ciclos y las menos fuertes (correspondientes a valores no tan bajos de las frecuencias) suelen corresponder a estaciones. Si hay dudas entre ciclos y estaciones podemos apoyarnos en las funciones de autocorrelación para discriminar.

- El ***periodograma acumulativo***, que representa en el eje de abscisas las frecuencias y en el de ordenadas las amplitudes acumuladas. Respecto del periodograma acumulativo tendremos en cuenta lo siguiente:
  - Para una serie aleatoria coincide con la diagonal del primer cuadrante.
  - Desvíos bruscos de la diagonal provocan presencia de ciclos o estaciones para las respectivas frecuencias, que serán ciclos cuando las frecuencias sean bajas.

La desestacionalización es una tarea no trivial que ha dado lugar a multitud de estudios y algoritmos, entre los que destacan los programas *X*11 y *X*12 del *Bureau of the Census* de Estados Unidos. A nivel trivial, existen varios métodos de desestacionalización. Los más sencillos son el método de la tendencia, el método de las medias móviles, el método de las diferencias estacionales y el método de las variables ficticias.

El *método de desestacionalización de las diferencias estacionales* permite eliminar la mayor parte del efecto estacional de una serie, y consiste en obtener la serie de diferencias de orden m (periodo estacional), definida como $z_t = y_t - y_{t-m}$.

De todos modos, es conveniente recordar que en cada diferenciación de orden m perdemos m observaciones de la serie original. *La decisión de diferenciar estacionalmente la serie* se basa en la FAC con el mismo criterio que para la diferenciación estacionaria pero considerando sólo los retardos referidos a periodos estacionales (m y sus múltiplos). Si los coeficientes de la FAC no decaen rápidamente en los retardos múltiplos del periodo estacional m hay que diferenciar estacionalmente la serie original.

La *eliminación de las variaciones estacionales, para inducir la estacionariedad*, suele hacerse casi siempre, mediante la diferenciación estacional. Si los datos son mensuales, la diferenciación estacional de la serie temporal X_t, consiste en calcular $Z_t = X_t - X_{t-12}$. Con datos trimestrales calcularíamos $Z_t = X_t - X_{t-4}$. Si después de efectuar esta transformación la serie sigue presentando evidencias de variaciones estacionales, es posible aplicar de nuevo el procedimiento, es decir, calcular las diferencias de segundo orden, y así sucesivamente.

MODELOS ESTACIONALES PUROS

Un modelo estacional de periodo s se denomina puro si sólo existe relación entre las observaciones que distan entre sí s periodos o múltiplos de s. En la práctica no serán éstos los modelos estacionales más habituales, sino que es común una estructura multiplicativa que mezcla la parte estacional con la parte regular estudiada en el capítulo anterior. De esta forma tenemos los modelos ARIMA generales.

Modelos autorregresivos estacionales AR(P)$_s$

Un modelo autorregresivo (AR) estacional describe una clase particular de proceso en el que las observaciones en un momento dado son predecibles a partir de las observaciones en los periodos estacionales previos del proceso más un término de error. El caso más simple es el AR (1,0,0)s, o AR(1)s o estacional de primer orden, cuya expresión matemática es:

$$X_t = \phi_1 X_{t-s} + a_t$$

El proceso autorregresivo estacional de orden p, representado por ARIMA(P,0,0)s, o simplemente por AR(P)s toma la forma:

$$X_t = \phi_1 X_{t-s} + \phi_2 X_{t-2s} + \ldots + \phi_P X_{t-Ps} + a_t$$

que puede ponerse, mediante el operador de cambio retroactivo B, en la forma:

$$(1 - \phi_1 B^s - \phi_2 B^{2s} - \ldots - \phi_P B^{Ps}) X_t = a_t \qquad B^k(X_t) = X_{t-k}$$

Un proceso autorregresivo estacional AR(P)s es *estacionario* si las raíces del polinomio en B dado por: $1 - \phi_1 B^s - \phi_2 B^{2s} - \ldots - \phi_P B^{Ps}$ caen fuera del círculo unidad.

La varianza de un proceso AR(1)s es: $g_0 = \dfrac{\sigma_a^2}{1-\phi_1}$

La función de autocovarianza de un proceso AR(1)s es:

$$g_k = \phi_1^k \frac{\sigma_a^2}{1-\phi_1} \quad k \geq 1$$

La función de autocorrelación de un proceso AR(1)s es :

$$h_k = \phi_1^k \quad k \geq 1$$

La función de autocorrelación parcial de un proceso AR(1)s es:

$$h_{kk} = \begin{cases} \phi_1 & para \quad j = s \\ 0 & para \quad j \neq s \end{cases}$$

La varianza de un proceso AR(2) es: $g_0 = \phi_1 g_s + \phi_2 g_{2s} + \sigma_a^2$

La función de autocovarianza de un proceso AR(2)s es:

$$g_k = \Phi_1 g_{(k-1)s} + \Phi_2 g_{(k-2)s} \quad k \geq 1$$

La función de autocorrelación de un proceso AR(2)s es :

$$h_k = \Phi_1 h_{(k-1)s} + \Phi_2 h_{(k-2)s} \quad k \geq 1$$

La función de autocorrelación parcial de un proceso AR(2)s es:

$$h_{kk} = \begin{cases} h_1 = \dfrac{\phi_1}{1-\phi_2} & para \quad j = s \\ \dfrac{h_2 - h_1^2}{1-h_1^2} = \phi_2 & para \quad j = 2s \\ 0 & para \quad j \neq s \text{ y } 2s \end{cases}$$

Modelos de medias móviles estacionales MA(Q)$_s$

En un modelo MA(Q)s el valor actual puede predrecirse a partir de la componente aleatoria de este momento y, en menor medida, de los impulsos aleatorios anteriores en los periodos múltiplos del periodo estacional. El modelo ARIMA(0,0,1)s, también denotado por MA(1)s, viene dado por la expresión:

$$X_t = a_t - \Theta_1 a_{t-s}$$

El proceso de medias móviles de orden Q, representado por ARIMA(0,0,Q)s, o también por MA(Q)s, viene dado por la expresión:

$$X_t = a_t - \Theta_1 a_{t-s} - \Theta_2 a_{t-2s} - .... - \Theta_Q a_{t-Qs}$$

que puede ponerse, mediante el operador de cambio retroactivo B, en la forma:

$$X_t = (1 - \Theta_1 B^s - \Theta_2 B^{2s} - .... - \Theta_Q B^{Qs})\, a_t$$

Un proceso de medias móviles es siempre estacionario.

Un proceso de medias móviles *MA(q)* es *invertible* si las raíces del polinomio en B definido por: $1 - \Theta_1 B^s - \Theta_2 B^{2s} - .... - \Theta_Q B^{Qs}$ caen fuera del círculo unidad.

La varianza de un proceso MA(1)s es: $g_0 = \sigma_a^2 (1 + \Theta_1^2)$

La función de autocovarianza de un proceso MA(1)s es:

$$g_k = \begin{cases} -\Theta_1 \sigma_a^2 & para \quad k = s \\ 0 & para \quad k \neq s \end{cases}$$

La función de autocorrelación de un proceso MA(1)s es :

$$h_k = \begin{cases} \dfrac{-\Theta_1}{1+\Theta_1^2} & para \quad k = s \\ 0 & para \quad k \neq s \end{cases}$$

La función de autocorrelación parcial de un proceso MA(1)s es:

$$h_{kk} = \frac{-\Theta_1^k (1-\Theta_1^2)}{1-\Theta_1^{2(k+1)}} \quad para \quad k \geq 1$$

La varianza de un proceso MA(2)s es: $g_0 = \sigma_a^2 (1 + \Theta_1^2 + \Theta_2^2)$

La función de autocovarianza de un proceso MA(2)s es:

$$g_k = \begin{cases} -(\Theta_1 + \Theta_1\Theta_2)\sigma_a^2 & para \quad k = s \\ -\Theta_2 \sigma_a^2 & para \quad k = 2s \\ 0 & para \quad k \neq 2 \text{ y } 2s \end{cases}$$

La función de autocorrelación de un proceso MA(2)s es:

$$h_k = \begin{cases} \dfrac{-\Theta_1 + \Theta_1\Theta_2}{1+\Theta_1^2+\Theta_2^2} & para \quad k=1 \\ \dfrac{-\Theta_2}{1+\Theta_1^2+\Theta_2^2} & para \quad k=2 \\ 0 & para \quad k>2 \end{cases}$$

La función de autocorrelación parcial de un proceso MA(2)s es:

$$h_{kk} = \begin{cases} h_s & para \quad k=s \\ \dfrac{h_{2s}-h_s^2}{1-h_s^2} & para \quad k=2s \\ \dfrac{h_s^3 - h_s h_{2s}(2-h_{2s})}{1-h_{2s}^2 - 2h_{1s}^2(1-h_{2s})} & para \quad k=3s \\ \ldots \end{cases}$$

Modelos estacionales ARMA(P,Q)$_s$

Una extensión natural de los *modelos AR(P)s* y *MA(Q)s* es un tipo de modelos que incluyen tanto términos autorregresivos estacionales como de medias móviles y se definen como ARMA(P,Q)s o también como ARIMA(P,0,Q)s. Se representan por la ecuación:

$$X_t = \phi_1 X_{t-s} + \phi_2 X_{t-2s} + \ldots + \phi_p X_{t-Ps} + a_t - \Theta_1 a_{t-s} - \Theta_2 a_{t-2s} - \ldots - \Theta_Q a_{t-Qs}$$

que puede ponerse de la forma:

$$X_t - \phi_1 X_{t-s} - \phi_2 X_{t-2s} - \ldots - \phi_p X_{t-Ps} = a_t - \Theta_1 a_{t-s} - \Theta_2 a_{t-2s} - \ldots - \Theta_Q a_{t-Qs}$$

o sea:

$$(1 - \phi_1 B^s - \phi_2 B^{2s} - \ldots - \phi_P B^{Ps})\, X_t = (1 - \Theta_1 B^s - \Theta_2 B^{2s} - \ldots - \Theta_Q B^{Qs})\, a_t$$

El proceso ARMA(P,Q)s es estacionario si lo es su componente autorregresiva, y es invertible si lo es su componente de medias móviles. Por tanto podemos decir que un modelo ARMA(P,Q)s es invertible si las raíces del polinomio en B definido mediante $1 - \Theta_1 B^s - \Theta_2 B^{2s} - \ldots - \Theta_Q B^{Qs}$ caen fuera del círculo unidad.

Un modelo ARMA(P,Q)s es estacionario si las raíces del polinomio definido por $1 - \phi_1 B^s - \phi_2 B^{2s} - \ldots - \phi_P B^{Ps}$ caen fuera del círculo unidad.

La varianza de un proceso ARMA(1,1)s es: $g_0 = \dfrac{\sigma_a^2(1+\Theta_1^2 - 2\phi_1\Theta_1)}{1-\phi_1^2}$

La función de autocovarianza de un proceso ARMA(1,1)s es:

$$g_k = \begin{cases} \dfrac{\sigma_a^2(1-\phi_1\Theta_1)(\phi_1-\Theta_1)}{1-\phi_1^2} & para \quad k=1 \\ \phi_1 g_{(k-1)s} & para \quad k>1 \end{cases}$$

La función de autocorrelación de un proceso ARMA(1,1) es:

$$h_k = \begin{cases} \dfrac{(1-\phi_1\Theta_1)(\phi_1-\Theta_1)}{1-\phi_1^2 - 2\phi_1\Theta_1} & para \quad k=1 \\ \phi_1 h_{(k-1)s} & para \quad k>1 \end{cases}$$

La función de autocorrelación parcial de un proceso ARMA(p,q) es:

$$h_{kk} = \begin{cases} h_s & para \quad k=s \\ \dfrac{h_{2s}-h_s^2}{1-h_1^2} & para \quad k=2s \\ \dfrac{h_s^3 - h_s h_{2s}(2-h_{2s}) + h_{3s}(1-h_s^2)}{1-h_{2s}^2 - 2h_s^2(1-h_{2s})} & para \quad k=3s \\ \ldots \end{cases}$$

Modelos ARIMA(P,D,Q)s estacionales puros

Un modelo ARIMA(0,D,0)s estacional puro es una serie temporal que se convierte en un ruido blanco (proceso puramente aleatorio) después de ser diferenciada D veces estacionalmente. El modelo general ARIMA(P,D,Q)s denominado proceso autorregresivo integrado de medias móviles de orden P, D, Q, toma la siguiente expresión:

$$(1 - \phi_1 B^s - \phi_2 B^{2s} - \ldots - \phi_P B^{Ps})(1-B^s)^D\, Y_t = (1 - \Theta_1 B^s - \Theta_2 B^{2s} - \ldots \Theta_Q B^{Qs})a_t$$

Un modelo ARIMA(P,D,Q)s estacional puro permite describir una serie de observaciones después de que hayan sido diferenciadas D veces, a fin de extraer las posibles fuentes de no estacionariedad. Esta fórmula general se puede aplicar a cualquier modelo. Si hay alguna componente P,D,Q igual a cero, se elimina el término correspondiente de la fórmula general.

Identificación de modelos estacionales puros

Para identificar modelos estacionales puros se siguen las mismas reglas que en los no estacionales, teniendo presente ahora que las gráficas de la funciones de autocorrelación y autocorrelación parcial son similares al caso no estacional pero para retardos múltiples del periodo estacional. Por ejemplo, en la Figura 6-9 se muestran las funciones de autocorrelación de un modelo AR(1) y de un modelo AR(1)s con estacionalidad trimestral. Se observa que las estructuras de las dos funciones de autocorrelación son semejantes considerando los retardos múltiples del periodo estacional en el caso de la serie estacional pura.

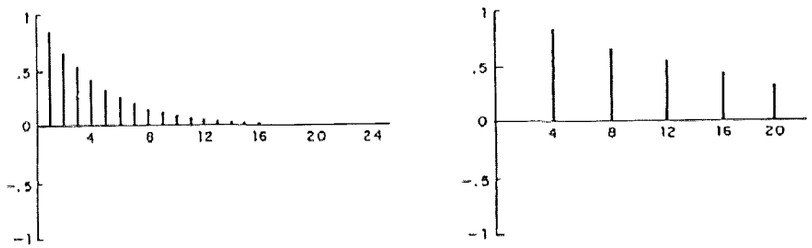


Figura 6-9

Para identificar un modelo AR(2)s basta observar que su función de autocorrelación se comporta como la de un AR(2) no estacional, pero considerando los retardos múltiplos del periodo estacional. La primera línea de dos gráficos de la Figura 6-10, o sea, Figuras a) y b), muestran funciones de autocorrelación de modelos AR(2)s. Este mismo criterio se sigue para identificar las estructuras MA(1)s, MA(2)s y ARMA(1,1)s. Las dos gráficas de la segunda línea de la Figura 3-2, o sea, Figuras c) y d), identifican modelos estacionales puros de medias móviles de orden 1 MA(1)s. Las dos gráficas de la tercera línea de la Figura 6-10, o sea, Figuras e) y f), identifican modelos estacionales puros de medias móviles de orden 2 MA(2)s. Las dos gráficas de la última línea de la Figura 6-10, o sea, Figuras g) y h), identifican modelos estacionales puros de ARMA(1,1)s. Se observa claramente que las estructuras de las funciones de autocorrelación son semejantes al caso no estacional considerando solamente los retardos de la función de autocorrelación relativos a los múltiplos del periodo estacional.

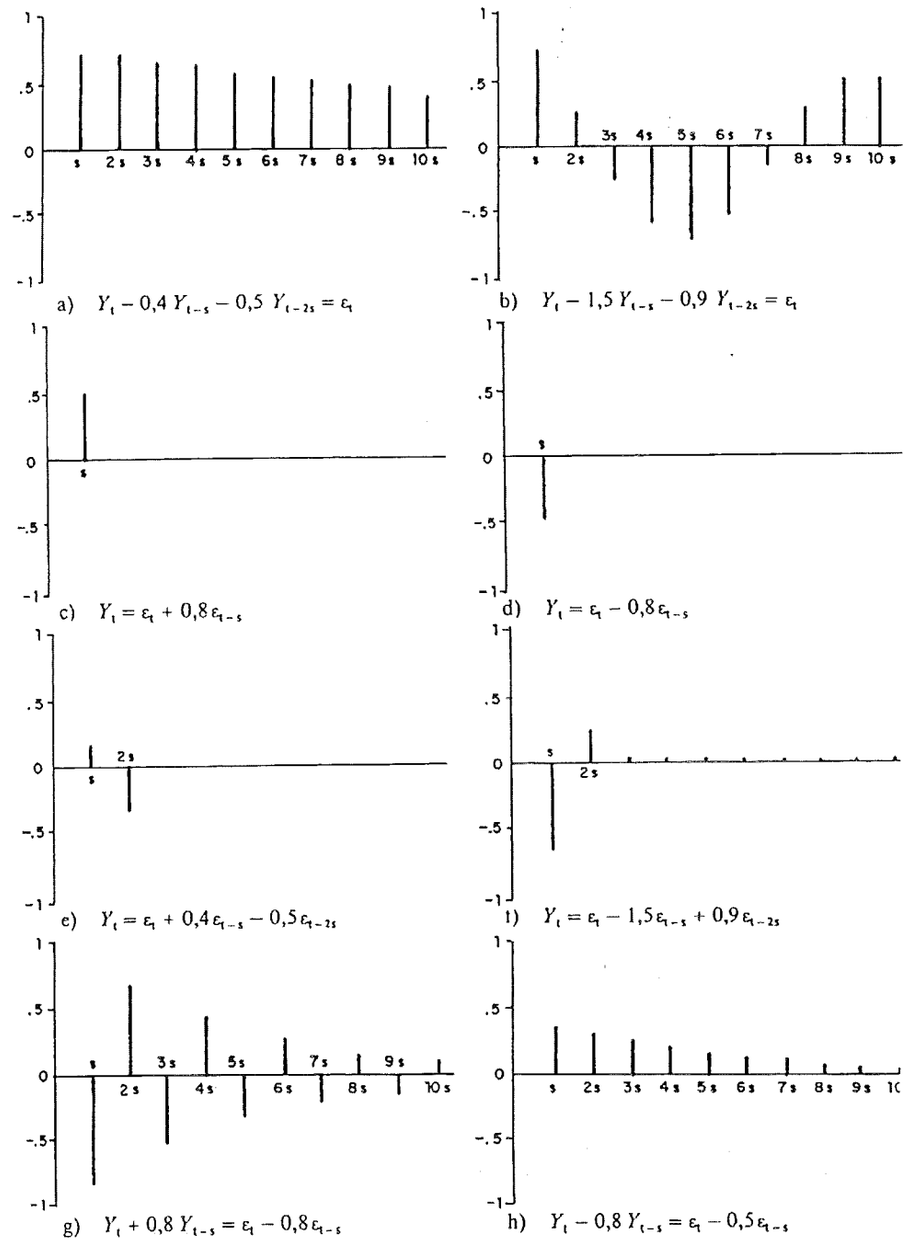


Figura 6-10

MODELOS ESTACIONALES GENERALES

En los modelos estacionales frecuentementemente no están solamente relacionadas las observaciones que distan entre sí múltiplos del periodo estacional, sino que lo habitual es que dentro de periodos no estacionales también existan relaciones. Los modelos que mezclan estos dos tipos de interrelaciones entre las observaciones son los modelos estacionales generales, también denominados modelos estacionales multiplicativos.

Un modelo estacional general será de la forma ARIMA$(p,d,q)(P,D,Q)s$ donde p, d y q son los parámetros de la parte regular y P,D y Q son los parámetros de la parte estacional. Su ecuación general podría expresarse en términos del operador diferencias B de la siguiente forma:

$$(1-\phi_1 B^s - \phi_2 B^{2s} -\ldots- \phi_P B^{Ps})\,(1- \Phi_1 B - \Phi_2 B^2 -\ldots- \Phi_P B^s)(1-B^s)^D\,(1-B)^d X_t =$$
$$(1 - \Theta_1 B^s - \Theta_2 B^{2s} - \ldots - \Theta_Q B^{Qs})\,(1- \nu_1 B - \nu_2 B^2 -\ldots- \nu_q B^q)a_t$$

Modelos estacionales generales con parte regular autorregresiva. Identificación

Consideraremos los casos más sencillos y más habituales en la práctica.

Para un modelo ***ARIMA(1,0,0)(1,0,0)$_{12}$*** la expresión será :

$$(1-\phi_1 B^{12})\,(1- \Phi_1 B)\,X_t = a_t$$

La características más importantes de su función de autocorrelación son :

1) Al ser la parte regular AR, las autocorrelaciones nunca se anulan.
2) Para retardos pequeños el comportamiento es el propio de la parte regular AR(1) siendo todos los coeficientes positivos si $\Phi_1>0$ y alternativos si $\Phi_1<0$.
3) En cuanto a los retardos estacionales, dado que la parte estacional es AR, las autocorrelaciones para estos retardos tampoco se anulan, siendo el comportamiento de la función de autocorrelación para ellos la propia de un proceso estacional de orden 1.
4) Las interdependencias entre las partes regular y estacional, finalmente se manifiestan claramente en los retardos próximos a los periodos estacionales.

La Figura 6-11 muestra distintas formas de las funciones de autocorrelación y autocorrelación parcial de un modelo ***ARIMA(1,0,0)(1,0,0)$_{12}$***.

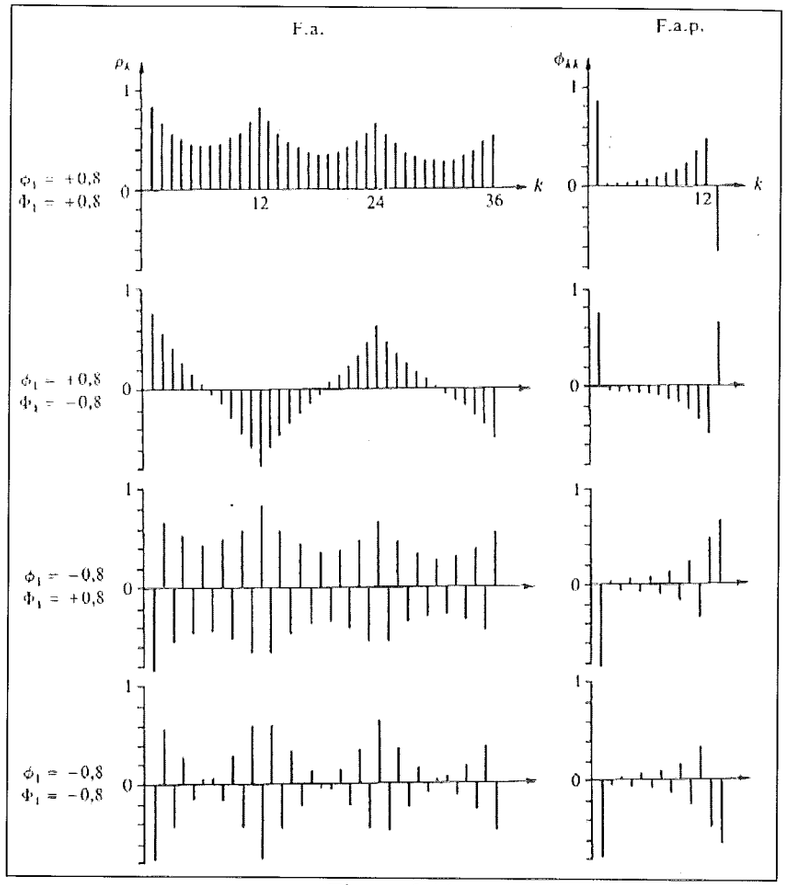


Figura 6-11

Para un modelo ***ARIMA(1,0,0)(0,0,1)$_{12}$*** la expresión será :

$$(1- \Phi_1 B)\,X_t = (1-\Theta_1 B^{12})\,a_t$$

En este tipo de modelos estacionales multiplicativos los comentarios son similares al caso anterior matizando que, como la parte estacional es ahora MA(1), sólo el coeficiente de autocorrelación estacional correspondiente al primer periodo estacional será claramente distinto de cero.

Por tanto la representación de la Figura 6-11 sigue siendo válida para modelos ***ARIMA(1,0,0)(0,0,1)$_{12}$*** con la salvedad de que sólo el coeficiente de autocorrelación estacional correspondiente al primer periodo estacional será claramente distinto de cero.

Para un modelo ***ARIMA(2,0,0)(1,0,0)$_{12}$*** la expresión será :

$$(1-\phi_1 B^{12})\,(1-\Phi_1 B-\Phi_2 B^2)\,X_t = a_t$$

La características más importantes de sus funciones de autocorrelación y autocorrelación parcial son:

1) La función de autocorrelación se amortigua lentamente siguiendo patrones de abanico con ciclo completo en el retardo 12 y sus múltiplos. La forma del patrón depende de los coeficientes del modelo.
2) La función de autocorrelación parcial tiene también un patrón estacional con salto en el retardo 12, correspondiente a la duración del periodo estacional en el modelo, y significación en los dos primeros retardos. Este comportamiento recuerda al de un AR(2) considerando los patrones de estacionalidad.

Las Figuras 6-12 y 6-13 muestran distintas formas de las funciones de autocorrelación y autocorrelación parcial de un modelo ***ARIMA(2,0,0)(1,0,0)$_{12}$***

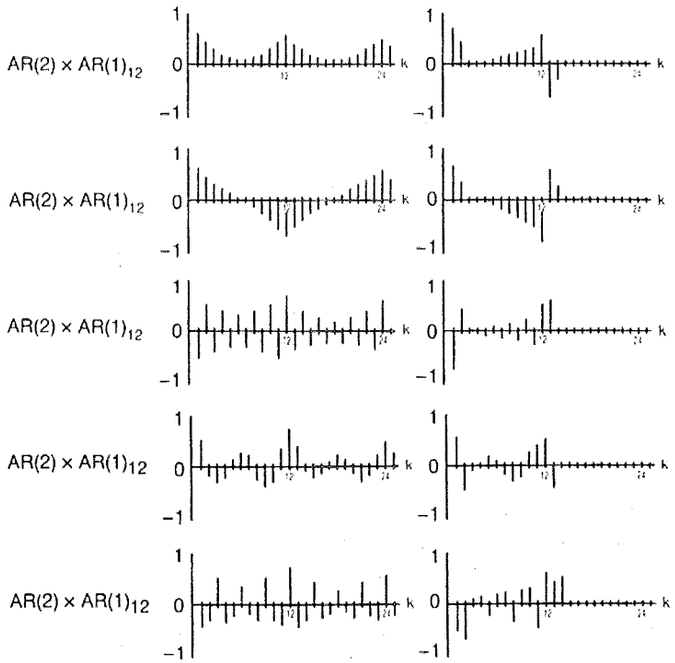


Figura 6-12

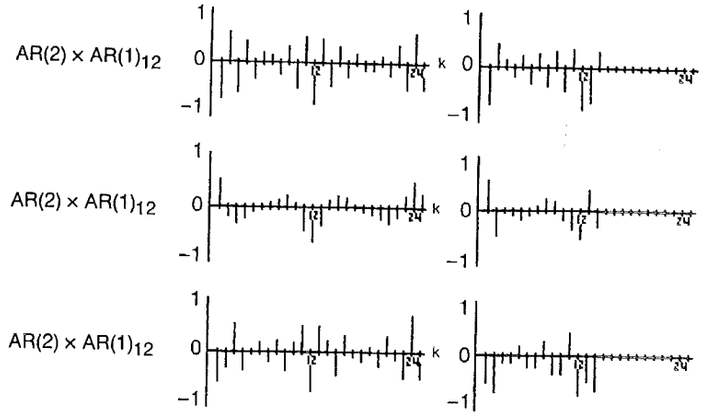


Figura 6-13

Modelos estacionales generales con parte regular de media móvil. Identificación

Para un modelo ***ARIMA(0,0,1)(0,0,1)$_{12}$*** , también denominado ***modelo líneas aéreas***, la expresión será:

$$X_t = (1-\Theta_1 B^{12})\,(1-\nu_1 B)\,a_t$$

Respecto de la función de autocorrelación tendremos en cuenta las siguientes características :

1) Dado que la parte estacional es MA(1)$_{12}$, las autocorrelaciones correspondientes a periodos estacionales sólo serán distintas de cero para el primer retardo estacional.

2) Dado que la parte regular del modelo es MA(1), el único valor del coeficiente de autocorrelación distinto de cero para retardos pequeños será el 1.

3) Las interdependencias entre las partes regular y estacional del modelo sólo afectarán a un periodo alrededor del retardo estacional distinto de cero, como consecuencia de que la parte regular del modelo es MA(1).

La Figura 6-14 muestra distintas formas de las funciones de autocorrelación (izquierda) y autocorrelación parcial (derecha) de un modelo ***ARIMA(0,0,1)(0,0,1)$_{12}$***.

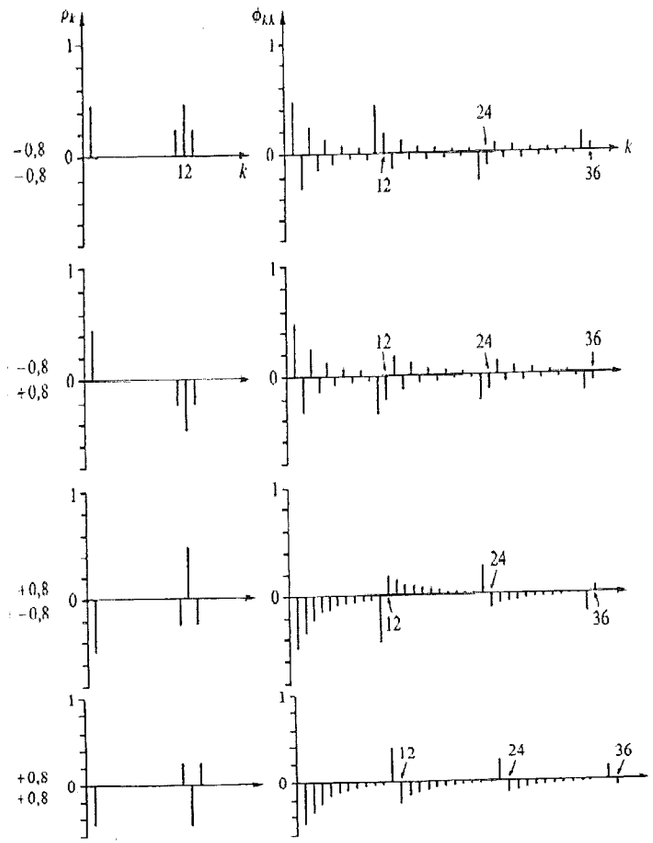


Figura 6-14

Para un modelo ***ARIMA(0,0,2)(0,0,1)$_{12}$*** la expresión será :

$$X_t = (1-\Theta_1 B^{12})\,(1-\nu_1 B-\nu_2 B^2)\,a_t$$

Respecto de la función de autocorrelación tendremos en cuenta las siguientes características :

1) Las consideraciones para la parte estacional son las mismas del caso anterior, ya que nuevamente estamos ante un MA(1)$_{12}$.

2) Dado que la parte regular del modelo es MA(2), los valores del coeficiente de autocorrelación distintos de cero para retardos pequeños serán 2.

3) Las interdependencias entre las partes regular y estacional del modelo afectan a dos periodos alrededor del retardo estacional distinto de cero, como consecuencia de que la parte regular del modelo es MA(2).

La Figura 6-15 muestra distintas formas de las funciones de autocorrelación (izquierda) y autocorrelación parcial (derecha) de un modelo ***ARIMA(0,0,2)(0,0,1)$_{12}$***

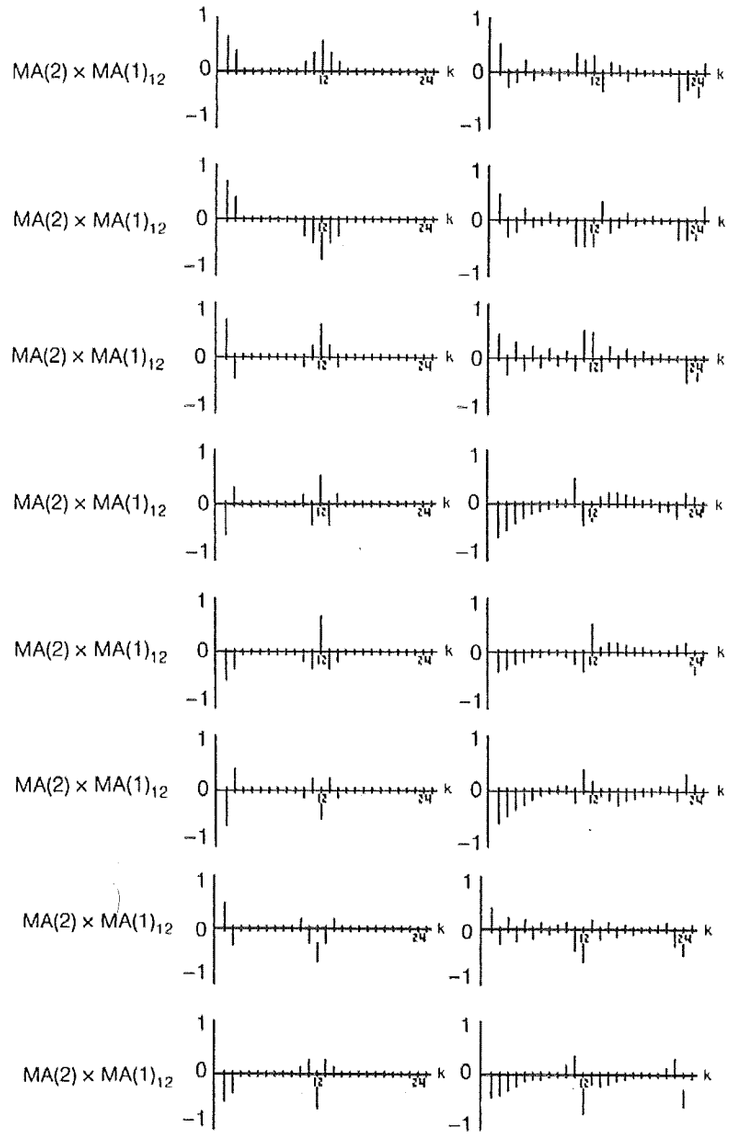


Figura 6-15

Para un modelo ***ARIMA(0,0,1)(1,0,0)$_{12}$*** la expresión será :

$$(1-\phi_1 B^{12})\, X_t = (1-\nu_1 B)\, a_t$$

Respecto de la función de autocorrelación tendremos en cuenta las siguientes características :

1) Dado que la parte estacional es AR(1), para los retardos estacionales múltiplos del periodo estacional, las autocorrelaciones no se anulan, si bien irán decreciendo progresivamente.

2) Dado que la parte regular del modelo es MA(1), los valores del coeficiente de autocorrelación distintos de cero para retardos pequeños serán únicamente el 1.

3) Las interdependencias entre las partes regular y estacional del modelo se concretan en que aparecerá un coeficinte de autocorrelación distinto de cero alrededor del retardo estacional, como consecuencia de que la parte regular del modelo es MA(1).

Las Figuras 6-16 y 6-17 muestran distintas formas de las funciones de autocorrelación de un modelo ***ARIMA(0,0,1)(1,0,0)$_{12}$***

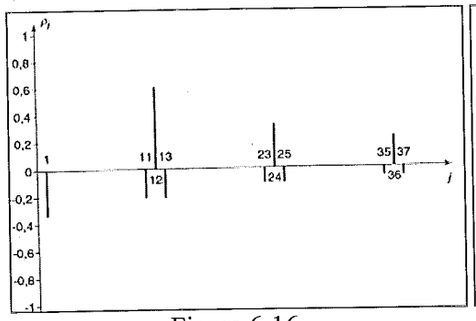


Figura 6-16

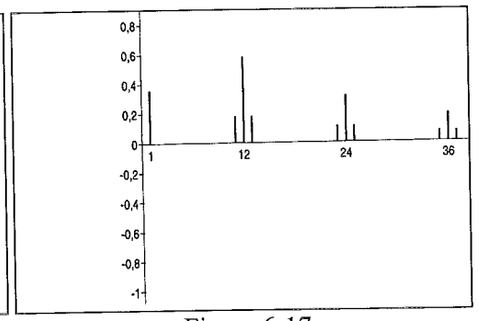


Figura 6-17

Para un modelo ***ARIMA(0,0,2)(1,0,0)$_{12}$*** la expresión será :

$$(1-\phi_1 B^{12})\, X_t = (1-\nu_1 B-\nu_2 B^2)\, a_t$$

Respecto de la función de autocorrelación tendremos en cuenta las siguientes características :

1) Las consideraciones para la parte estacional serán las mismas que en el caso anterior porque estamos nuevamente ante un AR(1)$_{12}$.

2) Dado que la parte regular del modelo es MA(2), los valores del coeficiente de autocorrelación distintos de cero para retardos pequeños serán 2.

3) Las interdependencias entre las partes regular y estacional del modelo afectan a dos periodos alrededor del retardo estacional distinto de cero, como consecuencia de que la parte regular del modelo es MA(2). Aparecerán dos coeficientes de autocorrelación distintos de cero alrededor del retardo estacional.

Las Figuras 6-18 a 6-21 muestran distintas formas de las funciones de autocorrelación de un modelo ***ARIMA(0,0,2)(1,0,0)$_{12}$***.

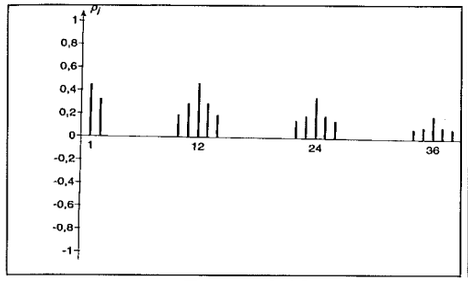


Figura 6-18

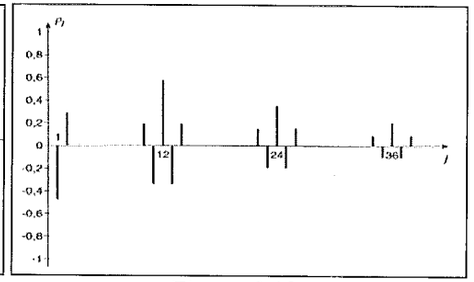


Figura 6-19

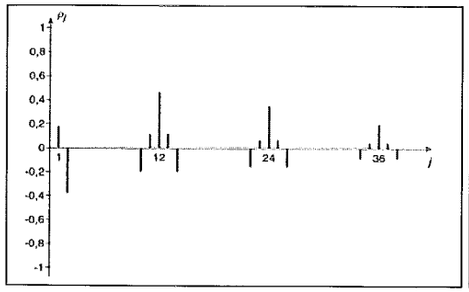


Figura 6-20

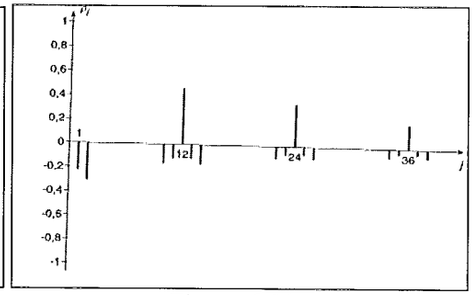


Figura 6-21

Identificación de modelos estacionales ARIMA(p,d,q)(P,D,Q)s

La primera tarea es siempre realizar diferenciaciones adecuadas estacionales y regulares que transformen la serie en estacionaria, es decir, identificar *D* y *d* para que las partes regular y estacional sean estacionarias. Para ver la estacionariedad de la parte estacional el método es el ya conocido para la parte regular pero considerando sólo retardos múltiples del periodo estacional, lo que exige bastantes valores en la serie. A continuación se analizarán las formas de las funciones FAC y FACP para adecuarlas a un determinado modelo ARMA. Para ello, el cuadro siguiente muestra pautas.

Modelo ARIMA	*Coeficientes de la FAC*	*Coeficientes de la FACP*
ARIMA(p,0,0)(P,00)s (gráficos g) y h) en 3-14)	Patrón estacional en *s*. Amortiguamiento lento	Pierden significación en el retardo *p*+1
ARIMA(0,0,q)(0,0,Q)s (gráficos a) y b) en 3-14)	Pierden significación en el retardo *q*+1	Patrón estacional en *s*-1. Amortiguamiento lento
ARIMA(p,0,q)(P,0,Q)s	Patrón estacional para la zona del retardo *s*. Amortiguamiento lento	Patrón estacional para la zona del retardo *s*. Amortiguamiento lento

La Figura 6-22 muestra las FAC de varios modelos estacionales generales. Los gráficos a) y b) se ajustan a un modelo de líneas aéreas ARIMA(0,0,1)(0,0,1)s. Los gráficos c) y d) se ajustan a un modelo ARIMA(0,0,2)(0,0,1)s. Los gráficos e) y f) se ajustan a un modelo ARIMA(0,0,1)(1,0,0)s. Los gráficos g) y h) se ajustan a un modelo ARIMA(1,0,0)(1,0,0)s.

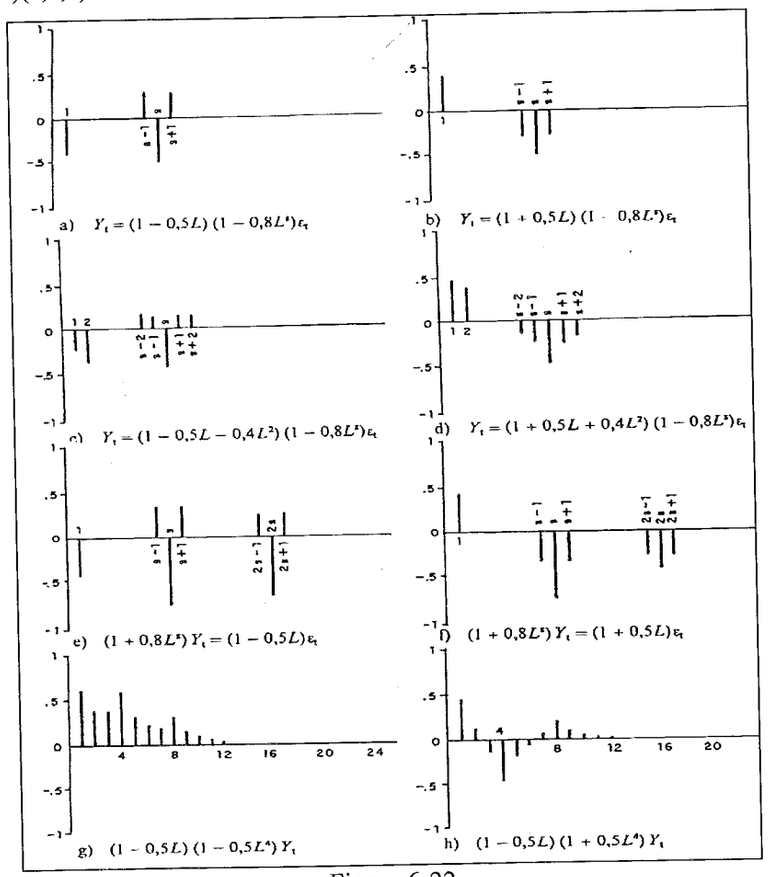


Figura 6-22

Estimación de modelos ARIMA(p,d,q)(P,D,Q)s

Considerando que un modelo ARMA(p,q)(P,Q)s puede escribirse siempre en su forma extendida como un modelo ARMA(Ps+p,Qs+q), el análisis de la estimación en los modelos ARMA regulares puede ser válido para extenderlo a los modelos ARMA estacionales generales.

Partiendo del modelo ARIMA(p,d,q) siguiente :

$$(1- \Phi_1 B - \Phi_2 B^2 -\ldots- \Phi_p B^s)\,(1-B)^d X_t = \delta+(1- \nu_1 B - \nu_2 B^2 -\ldots- \nu_q B^q)a_t$$

y considerando $w_t = (1-B)^d X_t$ tenemos:

$$w_t = \Phi_1 w_{t-1} - \Phi_2 w_{t-2} -\ldots- \Phi_p w_{t-p}+ \delta+ a_t - \nu_1 a_{t-1} - \nu_2 a_{t-2} -\ldots- \nu_q a_{t-q}$$

En el proceso de estimación se trata de obtener los mejores estimadores posibles para los parámetros del modelo $\beta' = (\delta, \Phi_1, \Phi_2, \ldots \Phi_p)$ y σ_a^2.

Los dos métodos de estimación más utilizados son el de mínimos cuadrados y el de máxima verosimilitud (que es el más recomendable). El método de máxima verosimilitud exige que los a_t sean normales idénticamente distribuidos $a_t \rightarrow \mathrm{NID}(0, \sigma_a^2)$ y que w_t sea un proceso estacionario e invertible.

Si aplicamos el *criterio de mínimos cuadrados*, el problema consiste en obtener β que minimice la función objetivo:

$$\sum_{t=1}^{T}[a_t(\beta / w)]^2 = \sum_{t=1}^{T}\hat{a}_t^2$$

La solución de este problema exige el conocimiento de valores iniciales:

$$\boldsymbol{w}^{0\prime}=(w_0, w_{-1}, \ldots, w_{1-p}) \quad \boldsymbol{u}^{0\prime}=(u_0, u_{-1},\ldots,u_{1-q})$$

Además, cuando el modelo tiene términos de medias móviles, no será lineal en los parámetros, por lo que deberán aplicarse métodos de estimación no lineales.

Conocidos los valores iniciales, podemos aplicar el ***enfoque condicional de estimación*** mediante el *método de máxima verosimilitud* (se llama enfoque condicional de estimación porque utilizada valores iniciales previamente conocidos). La función de verosimilitud condicional a los valores iniciales mencionados es:

$$L(\beta,\sigma_a^2 / w, w^0, u^0) = \left(2\pi\sigma_a^2\right)^{-T/2} \exp\left\{-\frac{1}{2\sigma_a^2}\sum_{t=1}^{T}[a_t(\beta / w, w^0, u^0)]^2\right\}$$

$$L(\beta,\sigma_a^2 / w, w^0, u^0) = \left(2\pi\sigma_a^2\right)^{-T/2} \exp\left\{-\frac{1}{2\sigma_a^2}\sum_{t=1}^{T}\hat{a}_t^2\right\}$$

En el proceso de estimación se trata de obtener los mejores estimadores posibles para los parámetros del modelo β y σ_a^2.

Otra alternativa es el ***enfoque no condicional de estimación*** en el que no se asume el conocimiento de valores iniciales y que parte de la *función de verosimilitud exacta*:

$$L(\beta,\sigma_a^2 / w, w^0, u^0) = \left(2\pi\sigma_a^2\right)^{-T/2} \exp\left\{-\frac{1}{2\sigma_a^2}\sum_{t=1-p-q}^{T}[E_c(a_t)]^2\right\}$$

donde $E_c(a_t)$ es la esperanza condicional de a_t a los valores muestrales de la serie w y de los parámetros β y σ_a^2.

Los estimadores máximo verosímiles exactos serán, por tanto, los valores de β y σ_a^2 que maximicen la función de verosimilitud exacta.

Validación de modelos ARIMA(p,d,q)(P,D,Q)s y predicción

Las tareas de validación de un modelo estacional general son equivalentes a las ya expresadas en el capítulo anterior para un modelo ARIMA. Pueden considerarse las siguientes fases :

Análisis de los coeficientes estimados. Estudio de la significatividad individual y conjunta de los coeficientes estimados.

Análisis de los residuos. Los residuos deben tener media cero, varianza constante, estar incorrelados y distribuirse normalmente. Su comportamiento ha de acercarse al de un ruido blanco. Los contrastes para estas tareas ya se vieron en el capítulo anterior.

Análisis de la estabilidad. Se trata de realizar contrastes para asegurar la validez del modelo en el futuro. Para ello se aplican los contrastes típicos de estabilidad estructural. Una prueba sencilla de estabilidad es tomar dos periodos muestrales distintos y estimar los parámetros para cada periodo. El modelo será estable si ambas estimaciones no son muy diferentes.

Una vez validado el modelo será útil para predecir. Las tareas de predicción son similares a las expuestas en el capítulo anterior. También podemos *evaluar su capacidad predictiva* mediante los contrastes adecuados.

MODELOS DE INTERVENCIÓN

La técnica del análisis de la intervención consiste en evaluar el efecto de determinadas intervenciones en el comportamiento de una serie temporal. Estas intervenciones se derivan del hecho de que las series reales se ven con frecuencia afectadas (intervenidas) por sucesos puntuales conocidos como huelgas, años bisiestos, cambios legales, accidentes, cambios en una festividades. Si realizamos una modelización de estos efectos en la serie podemos mejorar la precisión de la estimación de los parámetros y de las previsiones.

La primera idea para la incorporación de una intervención sería considerar una variable ficticia con valor uno en el periodo de la intervención y con valor cero en los restantes periodos, aunque la solución será un poco más compleja.

Box y Tiao (1975) denominaron análisis de intervención a la inclusión en un modelo de series temporales de variables ficticias para representar sucesos que producen efectos deterministas. Las variables ficticias más utilizadas para representar sucesos cualitativos que afectan a la serie son de dos tipos: *variables impulso* y *variables escalón*. Las variables impulso representan sucesos que ocurren únicamente en un instante, por ejemplo, un accidente, un error de medida o una huelga. Las variables escalón representan acontecimientos que comienzan en un instante conocido y se mantienen a partir de ese instante, por ejemplo, una subida de precios, un cambio legal, un cambio de base en un índice, etc.

En el análisis de la intervención el periodo de ocurrencia del suceso es conocido, pero desconocemos el periodo de comienzo de los mismos, aunque su existencia puede detectarse al observar datos atípicos en la serie (*outliers*). Habitualmente, cuando el tiempo y la ocurrencia de los sucesos externos de la intervención se desconoce, se dice que estamos trabajando con *outliers*.

Variables escalón e impulso

Una variable escalón se utiliza cuando en una serie temporal ocurre algún suceso que hace que los valores de esta serie, a partir del mismo en adelante, disminuyan o aumenten de forma continuada. La variable escalón (*step variable*) podría representarse como sigue:

$$S_t^{t_0} = \begin{cases} 0 & si \quad t < t_0 \\ 1 & si \quad t \geq t_0 \end{cases}$$

Una variable impulso se utiliza cuando en una serie temporal ocurre algún suceso que hace que un valor de la serie aumente o disminuya puntualmente y, a partir del mismo en adelante, la serie sigue la evolución que presentaba antes del suceso. La variable impulso (*pulse variable*) podría representarse como sigue:

$$P_t^{t_0} = \begin{cases} 0 & si \quad t \neq t_0 \\ 1 & si \quad t = t_0 \end{cases}$$

Considerando el operador en diferencias B, se observa que las variables impulso y escalón se relacionan mediante la exprersión:

$$P_t^{t_0} = (1-B)S_t^{t_0}$$

Cuando la duración del efecto del suceso de la intervención se considera permanente debemos utilizar variables escalón, mientras que cuando la duración del efecto es temporal se deben utilizar variables tipo impulso.

Comienzo brusco y duración permanente de una intervención

Cuando el impacto de una intervención es fijo aunque de magnitud desconocida ω y comienza en un periodo conocido t_0, el output de la intervención X_t puede escribirse como una variable escalón:

$$X_t = \omega S_t^{t_0}$$

Hemos supuesto que la intervención tiene impacto directo sobre el nivel de los datos, pero si lo tuviera sobre las variaciones de los datos, podríamos escribir:

$$(1-B)X_t = \omega S_t^{t_0}$$

Si la intervención surte efecto con un periodo de retraso podemos escribir:

$$X_t = \omega B S_t^{t_0}$$

Si el efecto de la intervención aparece después de b desfases temporales, el output de la intervención puede modelizarse como sigue:

$$X_t = \omega B^b S_t^{t_0}$$

Comienzo gradual y duración permanente de una intervención

Cuando la intervención tiene un comienzo gradual y una duración permanente podemos escribir:

$$X_t = \frac{\omega B^b}{1-\delta B} S_t^{t_0} \Rightarrow \begin{cases} X_t = \dfrac{\omega B}{1-\delta B} S_t^{t_0} & si\ b=1 \Rightarrow \begin{cases} X_t = \dfrac{\omega B}{1-B} S_t^{t_0} & si\ \delta = 1 \\ X_t = \omega B S_t^{t_0} & si\ \delta = 0 \end{cases} \\ X_t = \dfrac{\omega}{1-\delta B} S_t^{t_0} & si\ b = 0 \end{cases}$$

Cuando $b = 1$ y $\delta = 1$ el modelo lo podemos escribir como:

$$X_t = \frac{\omega B}{1-B} S_t^{t_0} \Rightarrow X_t = X_{t-1} + \omega S_{t-1}^{t_0}$$

Cuando $0 < \delta < 1$ el modelo lo podemos escribir como:

$$X_t = \delta X_{t-1} + \omega S_{t-1}^{t_0}$$

Comienzo brusco y duración temporal de una intervención

Cuando el impacto de una intervención tiene una duración temporal debemos considerar que el output de la intervención X_t puede escribirse como una variable impulso:

$$X_t = \frac{\omega B^b}{1-\delta B} P_t^{t_0} \Rightarrow \begin{cases} X_t = \dfrac{\omega B}{1-\delta B} P_t^{t_0} & si\ b=1 \Rightarrow \begin{cases} X_t = \dfrac{\omega B}{1-B} P_t^{t_0} & si\ \delta = 1 \\ X_t = \omega B P_t^{t_0} & si\ \delta = 0 \end{cases} \\ X_t = \dfrac{\omega}{1-\delta B} P_t^{t_0} & si\ b = 0 \end{cases}$$

Cuando $b = 1$ y $\delta = 1$ el modelo lo podemos escribir como:

$$X_t = \frac{\omega B}{1-B} P_t^{t_0} \Rightarrow X_t = X_{t-1} + \omega P_{t-1}^{t_0}$$

Cuando $0<\delta<1$ el modelo lo podemos escribir como:

$$X_t = \delta X_{t-1} + \omega P_{t-1}^{t_0}$$

Comienzo gradual y duración temporal de una intervención

Cuando el impacto de una intervención aumenta gradualmente hasta alcanzar un nivel máximo a partir del cual va disminuyendo, el output de la intervención X_t se especificará mediante modelos más complejos, como por ejemplo:

$$X_t = \frac{\omega B}{1-\delta_1 B-\delta_2 B^2} P_t^{t_0}$$

En el caso $b = 0$ podríamos escribir:

$$X_t = \frac{\omega}{1-\delta_1 B-\delta_2 B^2} P_t^{t_0} \Rightarrow X_t = \delta_1 X_{t-1} + \delta_2 X_{t-2} + \omega P_t^{t_0}$$

Combinación de variables escalón e impulso

Cuando se combinan variables escalón e impulso, el modelo de intervención puede escribirse como sigue:

$$X_t = \left(\frac{\omega_1 B}{1-\delta B} + \frac{\omega_2 B}{1-B}\right) P_t^{t_0} \Rightarrow$$

$$\Rightarrow X_t = (1+\delta)X_{t-1} - \delta X_{t-2} + (\omega_1+\omega_2)P_{t-1}^{t_0} - (\omega_1+\omega_2\delta)P_{t-2}^{t_0}$$

Modelo de intervención general

Cualquier modelo de intervención puede especificarse mediante la expresión:

$$X_t = \nu(B) I_t^{t_0}$$

siendo $I_t^{t_0}$ una variable impulso o escalón y además:

$$\nu(B) = \frac{\omega(B)B^b}{\delta(B)}$$

$$\omega(B) = \omega_0 - \omega_1 B - \omega_2 B^2 - \omega_s B^s$$
$$\delta(B) = \delta_0 - \delta_1 B - \delta_2 B^2 - \delta_r B^r$$

Combinado valores de b entre cero y uno y de r entre cero y dos, se obtienen como casos particulares todos los modelos estudiados anteriormente.

Se puede generalizar el modelo simple anterior a k intervenciones como sigue:

$$X_t = \sum_{j=1}^{k} \nu_j(B) I_t^{t_0 j}$$

$$\nu_j(B) = \frac{\omega_j(B)B^{bj}}{\delta_j(B)}$$

La Figura 6-23 representa los siguientes elementos:

- En su primera columna los órdenes r y s de los polinomios $\omega(B)$ y $\delta(B)$.
- En su segunda columna la ecuación del modelo de intervención adecuado.
- En la tercera columna la función escalón asociada $(x_{\mathfrak{I}} = 1, t \geq \mathfrak{I})$.
- En la cuarta columna presenta la variable impulso asociada $(x_{\mathfrak{I}} = 1, t = \mathfrak{I})$.

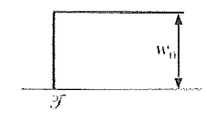
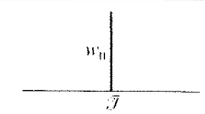
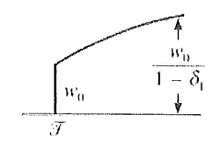
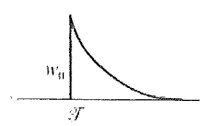
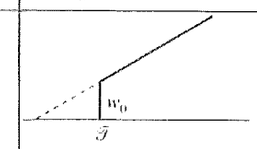
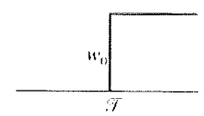
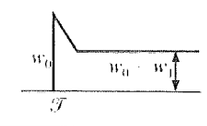
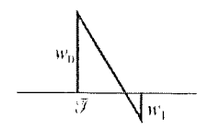
$r = 0$ $s = 0$	$y_t = w_0 x_t + n_t$		
$r = 1$ $s = 0$	$y_t = \frac{w_0}{(1-\delta_1 B)} x_t + n_t$		
$r_1 = 1$ $s = 0$ $\delta_1 = 1$	$y_t = \frac{w_0}{1-B} x_t + n_t$		
$r = 0$ $s = 1$	$y_t = (w_0 - w_1 B)x_t + n_t$		

Figura 6-23

IDENTIFICACIÓN DE MODELOS DE INTERVENCIÓN

La finalidad ahora es poder identificar modelos ARIMA con intervenciones. Para ello partimos de la expresión general del modelo ARIMA asociado a la serie temporal y_t:

$$\phi(B)(1-B)^d y_t = \theta(B)u_t$$

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$
$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$$

El modelo anterior puede escribirse también como:

$$y_t = \frac{\theta(B)}{\phi(B)(1-L)^d} u_t$$

Para modelos ARIMA(p,d,q)(P,D,Q)s estacionales generales con periodo estacional s podemos escribir:

$$y_t = \frac{\Theta(B^s)\theta(B)}{\Phi(B^s)\phi(B)(1-B^s)^D(1-B)^d} u_t$$

$$\Phi(B) = 1 - \Phi_1 B^s - \cdots - \Phi_Q B^{Qs}$$
$$\Theta(B) = 1 - \Phi_1 B - \cdots - \Phi_P B^{Ps}$$

Si incorporamos la intervención a nuestro modelo tenemos:

$$y_t = \nu(B) I_t^{t_0} + \frac{\Theta(B^s)\theta(B)}{\Phi(B^s)\phi(B)(1-B^s)^D(1-B)^d} u_t = \nu(B) I_t^{t_0} + N_t$$

con:

$$N_t = \frac{\Theta(B^s)\theta(B)}{\Phi(B^s)\phi(B)(1-B^s)^D(1-B)^d} u_t$$

Generalizando para k intervenciones, tenemos:

$$y_t = \sum_{j=1}^{k} \nu_j(B) I_t^{t_0 j} + N_t$$

La tarea es ahora identificar este último modelo, para lo que no será ya válido utilizar las funciones de autocorrelación y autocorrelación parcial, como en el caso ARIMA.

Si el subconjunto de datos anteriores o posteriores a las intervenciones es suficientemente grande, podríamos identificar el modelo ARIMA del modo tradicional tomando como base de datos dicho subconjunto, incorporando posteriormente la modelización de las intervenciones.

Una vez identificado el modelo se estimarán conjuntamente todos los parámetros del mismo y se aplicarán todos los contrastes de validación o diagnosis.

Cuando no disponemos de suficientes datos antes o después de las intervenciones, o bien al observar la serie existe alguna intuición respecto a la forma en que las intervenciones han afectado al patrón de comportamiento de la serie, podemos especificar la forma de $\nu_j(B)$ y estimar este modelo, para obtener:

$$y_t = \sum_{j=1}^{k} \hat{\nu}_j(B) I_t^{t_0 j}$$

y a continuación hallamos los residuos mediante la expresión:

$$N_t = y_t - \sum_{j=1}^{k} \hat{\nu}_j(B) I_t^{t_0 j}$$

que podrían interpretarse como los valores de la serie original limpia de intervenciones. Estos residuos se utilizarán para identificar el modelo ARIMA univariante que siguen a través de las funciones de autocorrelación y autocorrelación parcial. A continuación combinamos el modelo de intervención y el modelo residual para obtener el modelo de intervención global que debe ser estimado conjuntamente por máxima verosimilitud. A continuación sería necesario validar el modelo esencialmente mediante el análisis de los residuos.

VALORES ATÍPICOS (*OUTLIERS*)

Cuando se desconoce el tiempo y la causa de los factores extremos creadores de observaciones atípicas en una serie temporal, estamos ante el *análisis de outliers*.

Es habitual que ocurran con mucha frecuencia en las series reales hechos puntuales que desconocemos, como por ejemplo que la serie puede haber estado sometida a cambios de base o que haya habido errores de medición, etc. Las observaciones afectadas por estas intervenciones pueden presentar una estructura distinta de las demás y aparecer como datos atípicos, es decir, datos que aparentemente no han sido generados igual que las demás.

Es fundamental identificar estas situaciones desconocidas y separarlas de la dinámica habitual de la serie porque si sus efectos son grandes, pueden sesgar la estimación de los parámetros, lo que producirá malas predicciones futuras. Además, si el suceso ha ocurrido en la última parte de la serie y alguna observación afectada se utiliza para generar predicciones éstas no serán buenas, incluso aunque los parámetros estén bien estimados. Por otro lado, si estos sucesos atípicos pueden volver a aparecer en el futuro y los identificamos y estimamos sus efectos, podemos incorporar esta información en las predicciones y obtener intervalos de predicción mejores.

Tipos de outliers

Un modelo ARIMA(p,d,q)(P,D,Q) puede expresarse como sigue:

$$z_t = \frac{\Theta(B^s)\theta(B)}{\Phi(B^s)\phi(B)(1-B^s)^D(1-B)^d}u_t$$

$$\phi(B) = 1-\phi_1 B - \cdots - \phi_p B^p$$
$$\theta(B) = 1-\theta_1 B - \cdots - \theta_q B^q$$
$$\Phi(B) = 1-\Phi_1 B^s - \cdots - \Phi_Q B^{Qs}$$
$$\Theta(B) = 1-\Phi_1 B - \cdots - \Phi_P B^{Ps}$$

En el caso particular de una ARMA(p,q) tendríamos:

$$z_t = \frac{\theta(B)}{\phi(B)}u_t$$

que es el modelo que consideraremos en lo que sigue.

Outliers aditivos (AO)

Un outlier aditivo es un suceso que afecta a la serie en un solo instante temporal $t=t_0$. Por tanto podrá expresarse como:

$$y_t = z_t + \omega I_t^{t_0} \quad \text{donde } I_t^{t_0} = \begin{cases} 0 & si \quad t \neq t_0 \\ 1 & si \quad t = t_0 \end{cases}$$

$I_t^{t_0}$ es la variable que representa la presencia del outlier en el periodo $t=t_0$ y ω es el efecto del citado outlier.

Para modelos ARMA(p,q) la presencia de un outlier aditivo se modeliza como:

$$y_t = \frac{\theta(B)}{\phi(B)}u_t + \omega I_t^{t_0}$$

Outliers innovacionales (IO)

Un outlier innovacional es un suceso cuyo efecto se propaga en conformidad con el modelo ARIMA del proceso, afectando a todos los valores observados después de la ocurrencia. Se puede representar como sigue:

$$y_t = z_t + \frac{\theta(B)}{\phi(B)}\omega I_t^{t_0} \quad \text{donde } I_t^{t_0} = \begin{cases} 0 & si \quad t \neq t_0 \\ 1 & si \quad t = t_0 \end{cases}$$

Esta expresión también puede escribirse como:

$$y_t = \frac{\theta(B)}{\phi(B)}u_t + \frac{\theta(B)}{\phi(B)}\omega I_t^{t_0} = \frac{\theta(B)}{\phi(B)}(u_t + \omega I_t^{t_0})$$

Outliers de cambio en nivel (LS)

Un cambio en nivel es un suceso que afecta a la serie en un periodo dado y cuyo efecto es permanente. Se puede representar como sigue:

$$y_t = z_t + \frac{1}{1-B}\omega I_t^{t_0} \quad \text{donde } I_t^{t_0} = \begin{cases} 0 & si \quad t \neq t_0 \\ 1 & si \quad t = t_0 \end{cases}$$

Esta expresión también puede escribirse como:

$$y_t = \frac{\theta(B)}{\phi(B)}u_t + \frac{\omega}{1-B}I_t^{t_0} = \frac{\theta(B)}{\phi(B)}(u_t + \omega I_t^{t_0})$$

También podemos escribir:

$$y_t = z_t + \omega S_t^{t_0} \quad \text{donde } S_t^{t_0} = \begin{cases} 0 & si \quad t < t_0 \\ 1 & si \quad t \geq t_0 \end{cases}$$

Podemos decir entonces que el modelo de un outlier aditivo AO y el de un cambio en nivel LS es el mismo pero dependiendo el outlier AO de una variable impulso, mientras que el outlier LS depende de una variable escalón.

Outliers de cambio temporal (TC)

Un cambio temporal es un suceso que tiene un impacto inicial y cuyo efecto decae exponencialmente en conformidad con un factor de amortiguación δ. Se puede representar como sigue:

$$y_t = z_t + \frac{1}{1-\delta B}\omega I_t^{t_0} \quad \text{donde } I_t^{t_0} = \begin{cases} 0 & si \quad t \neq t_0 \\ 1 & si \quad t = t_0 \end{cases} \qquad 0<\delta<1$$

Esta expresión también puede escribirse como:

$$y_t = \frac{\theta(B)}{\phi(B)}u_t + \frac{\omega}{1-\delta B}I_t^{t_0} = \frac{\theta(B)}{\phi(B)}(u_t + \omega I_t^{t_0}) \qquad 0<\delta<1$$

Como una serie temporal puede tener varios outliers de diferentes tipos, un modelo general de serie con outliers podría expresarse como sigue:

$$y_t = \sum_{j=1}^{k} \omega_j \nu_j(B) I_{jt} + z_t$$

$$z_t = \frac{\theta(B)}{\phi(B)}u_t \qquad \nu_j(B) = \begin{cases} 1 & para\ AO \\ \dfrac{\theta(B)}{\phi(B)} & para\ IO \\ \dfrac{1}{1-B} & para\ LS \\ \dfrac{1}{1-\delta B} & para\ TC \end{cases}$$

Los efectos de los outliers son independientes de la estructura ARIMA de las series salvo en el caso IO. Los outliers AO y LS son casos límite de TC que se obtienen para δ = 0 y δ = 1 respectivamente. El outlier AO causa un efecto inmediato y único en la serie observada en $t = t_0$ de magnitud ω. El outlier TC produce un efecto inicial ω en $t=t_0$ de modo que este efecto decae gradualmente en el tiempo con un factor de amortiguación δ. El outlier LS introduce un cambio brusco y permanece en la serie observada de tipo escalón. El outlier IO depende de la estructura ARIMA de la serie y es el más complicado.

Para detectar outliers existen procedimientos iterativos como el de Chang y Tiao (1983) y el de Tsay (1986) que detectan el outlier, lo asocian a un modelo de intervención de acuerdo con su tipo y finalmente estiman dicho modelo.

Más recientemente Chen y Liu (1990) han diseñado un procedimiento consistente en la detección de los outliers y la estimación conjunta de los parámetros del modelo y los efectos de los outliers.

MODELO UNIVARIANTE DE LA FUNCIÓN DE TRANSFERENCIA

En el enfoque propuesto por Box y Jenkins sobre modelos estocásticos de series temporales se suelen diferenciar cinco clases de modelos con un grado creciente de complejidad: son modelos univariantes, modelos de intervención, modelos de funciones de transferencia, modelos estocásticos multivariantes y modelos de funciones de transferencia multivariantes. Podríamos añadir también los modelos vectoriales autorregresivos VAR.

Dada una variable z_t el modelo univariante puede plantearse como:

$$\phi(B)z_t = \theta(B)a_t$$

$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ = operador polinomial de retardos indicativo de un proceso autorregresivo de orden p sobre la variable z_t considerada.

$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ = operador polinomial de retardos para un proceso de medias móviles de orden q definido sobre un término de error a_t que cumple las hipótesis tradicionales de media nula, varianza constante y ausencia de autocorrelación.

Al despejar z_t el modelo univariante puede expresarse como:

$$z_t = \frac{\theta(B)}{\phi(B)}a_t$$

Puede interpretarse como z_t como el output o salida producidos por un filtro lineal cuyo input o entrada es una variable aleatoria de error o ruido a_t con media nula, varianza constante y ausencia de autocorrelación (un ruido blanco).

La función de transferencia del filtro es entonces el cociente de los dos polinomiales de retardos definidos $\theta(B)/\phi(B)$.

Cuando estemos interesados en relacionar varias series temporales ya no son de aplicación los métodos estudiados hasta ahora para modelos univariantes. Será necesario construir un modelo multivariante de series temporales.

Supongamos que la variable de output y_t queda explicada por la unión de dos componentes, uno de error o de ruido e_t, y otro, y_t^*, que puede ser explicado exactamente en términos de una variable explicativa x_t, esto es, $y_t = y_t^* + e_t$ cumpliéndose:

$$y_t^* = \frac{\omega(B)}{\delta(B)} x_t$$

$$\delta(B) = 1 - \delta_1 B - \cdots - \delta_r B^r$$
$$\omega(B) = \omega - \omega_1 B - \cdots - \omega_s B^s$$

$$e_t = \frac{\theta(B)}{\phi(B)} a_t$$

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$
$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$$

Entonces puede definirse el modelo de función de transferencia simple, con ruido, como:

$$y_t = y_t^* + e_t \Rightarrow y_t = \frac{\omega(B)}{\delta(B)} x_t + \frac{\theta(B)}{\phi(B)} a_t$$

donde y_t puede interpretarse, a su vez, como la salida de dos filtros lineales, uno sobre la variable de input x_t y otro sobre la componente de ruido.

La generalización al caso de más de una variable explicativa es inmediata, con k cocientes distintos de polinomiales de retardos, uno por cada input:

$$y_t = \sum_{j=1}^{k} \frac{\omega_j(B)}{\delta_j(B)} x_{jt} + \frac{\theta(B)}{\phi(B)} a_t$$

En el caso particular de que las variables x_{ht} sean ficticias (en el sentido econométrico de variables con valores sólo cero y uno) el modelo de función de transferencia resulta ser un modelo de intervención. Por tanto, el modelo de intervención es un caso particular de modelo de la función de transferencia.

En el caso de que todos los polinomiales de retardos sean de orden cero, el modelo de la función de transferencia coincide con el modelo básico de regresión:

$$y_t = \sum_{j=1}^{k} \omega_j x_{jt} + a_t$$

Si fijamos un proceso autorregresivo cualquiera de orden p para el término de error y mantenemos los demás polinomiales en el orden cero, el modelo de la función de transferencia pasa a ser el modelo de regresión generalizado AR(p):

$$y_t = \sum_{j=1}^{k} \omega_j x_{jt} + e_t \quad \text{con} \quad e_t = \phi_1 e_{t-1} + \cdots + \phi_p e_{t-p} + a_t$$

En el caso de que el polinomio autorregresivo $\delta(B)$ sea de orden 1 y coincida con el correspondiente al término de error, $\phi(B) = \delta(B)$, y los demás polinomios sean nulos, estaremos ante un modelo de regresión con una estructura de retardos de Koyck que se estudiará en un capítulo posterior:

$$y_t = \delta_1 y_{t-1} + \sum_{j=1}^{k} \omega_j x_{jt} + (e_t - \delta_1 e_{t-1})$$

Resumiendo podemos decir que ***un modelo de la función de transferencia*** es un modelo multivariante de series temporales donde adicionalmente a la componente de ruido a_t existen múltiples inputs x_{jt} que tratan de explicar el comportamiento de la serie que se quiere obtener y_t mediante la ecuación:

$$y_t = \sum_{j=1}^{k} \frac{\omega_j(B)}{\delta_j(B)} x_{jt} + \frac{\theta(B)}{\phi(B)} a_t$$

Identificación, estimación y validación del modelo de la función de transferencia

Para exponer los conceptos básicos para la identificación del modelo de función de transferencia, partimos del caso mas simple de variables sin transformación (o previamente transformadas), sin componente estacional e incluso de una sola variable explicativa x_t, es decir, consideramos el modelo de la función de transferencia en su expresión más simple:

$$y_t = \frac{w(B)}{\delta(B)} x_t + \frac{\theta(B)}{\phi(B)} a_t$$

A efectos de identificar los órdenes de los polinomiales de retardos implicados, es preciso ahora disponer de la información adicional que nos suministra la ***función de correlación cruzada***, que para dos procesos estocásticos estacionarios, x e y, quedará definida por:

$$\rho_{xy}(k) = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y} \quad k = 0, \pm 1, \pm 2, \ldots \quad \gamma_{xy}(k) = E[(x_t - \mu_x)(y_{t+k} - \mu_y)]$$

$$\gamma_{xy}(k) = \gamma_{xy}(-k)$$

En la práctica, al igual que ocurría con la función de autocorrelación, dispondremos de las estimaciones muestrales de la función de correlación cruzada:

$$\hat{\rho}_{xy}(k) = r_{ry}(k) = \frac{C_{xy}(k)}{S_x S_y}$$

$$C_{xy}(k) = \frac{1}{N} \sum_{t=1}^{N-k} (x_t - \bar{x})(y_{t+k} - \bar{y})$$

S_x y S_y son las desviaciones típicas muestrales de las series x e y.

Expresado el modelo en función de los infinitos valores anteriores de la variable explicativa (la denominada función de respuesta o de impulso-respuesta) más el componente de error tenemos:

$$y_t = v_0 x_t + v_1 x_{t-1} + \cdots + n_t$$

Si multiplicamos por x_{t-k} y tomamos esperanzas matemáticas ya en desviaciones a la media, supuesto $\mu_x = \mu_y = 0$, tenemos:

$$\gamma_{xy}(k) = v_0 \gamma_{xx}(k) + v_1 \gamma_{xx}(k-1) + \cdots$$

Hemos eludido el término $\gamma_{xn}(k)$ por admitir por planteamiento la incorrelación entre x_t y n_t. Evidentemente, en el caso de que el proceso x_t tuviera autocovarianzas nulas, la anterior relación se reduciría a:

$$\gamma_{xy}(k) = v_k \gamma_{xx}(0) = v_k \sigma_x^2$$

$$\rho_{xy}(k) = \frac{v_k \sigma_x^2}{\sigma_x \sigma_y} = v_k \frac{\sigma_x}{\sigma_y}$$

Podemos obtener entonces una primera estimación de los coeficientes del modelo en forma de distribución infinita de retardos, a partir de los valores muestrales de la covarianza cruzada:

$$\hat{v}_k = \frac{C_{xy}(k)}{S_x^2}$$

o, en forma equivalente, del coeficiente de correlación cruzada:

$$\hat{v}_k = r_{xy}(k) \frac{S_y}{S_x}$$

A continuación analizamos cómo puede conseguirse una situación tal que la serie explicativa tenga autocovarianzas nulas. Para ello partamos del modelo univariante que permite transformar la serie x_t en ruido blanco α_t:

$$\phi_x(B) x_t = \theta_x(B) \alpha_t$$

que podemos expresar en forma de ponderaciones "psi":

$$x_t = \phi_x^{-1}(B) \theta_x(B) \alpha_t = \Psi^{-1}(B) x_t$$

Si ahora aplicamos ese operador $\Psi^1(B)$ a la serie modelizada y_t, operación que se denomina ***preblanqueado de la serie***, la nueva variable y_t^* puede expresarse como:

$$y_t^* = \Psi^{-1}(B) y_t = \frac{w(B)}{\delta(B)} \Psi^{-1}(B) x_t + \frac{\theta(B)}{\phi(B)} \Psi^{-1}(B) a_t = \frac{w(B)}{\delta(B)} \alpha_t + \Psi^{-1}(B) n_t = v_t(B) \alpha_t + n_t^*$$

En resumen, la serie preblanqueada y_t^* puede expresarse en términos de dos componentes de ruido blanco α_t y a_t, o bien en un modelo con un término estocástico de error también preblanqueado y una función de transferencia entre y_t^* y el ruido blanco α_t, igual a la originariamente existente entre y_t y x_t.

Como consecuencia, en el modelo $y_t^* = v_0 \alpha_t + v_1 \alpha_{t-1} + \cdots + n_t^*$ podemos obtener unas estimaciones iniciales de los parámetros a partir de:

$$\hat{v}_k = r_{\alpha y^*}(k)\frac{S_{y^*}}{S_\alpha}$$

Estas estimaciones no son eficientes, pero nos permiten la determinación de los órdenes de los polinomios w(B) y δ(B) de la función de transferencia del modelo original.

En cuanto a la especificación de la componente de error del modelo, suele ser práctica habitual el determinar los órdenes de los polinomios $\theta(B)$ y $\phi(B)$ aceptando como tales los que corresponden al modelo univariante de y_t. Otra alternativa es analizar el término de error estimado a partir de $n_t = y_t - v(B)x_t$, cortando en un valor arbitrario de v_i. No obstante, los residuos así obtenidos parten de estimaciones ineficientes de los coeficientes de la función de respuesta.

La generalización al caso de k variables explicativas es relativamente inmediata si las x_{jt} están aproximadamente incorrelacionadas entre sí. En esta situación, el preblanqueo se realizará para cada variable en forma aislada, tomando como estructura del término de error n_t la más compleja de todas las que corresponden a las diferentes funciones de transferencia de cada variable explicativa. Por tanto, para cada variable se procede a estimar su función de respuesta $v_j(B)$ en forma aislada, con lo que dispondremos de las orientaciones iniciales sobre los órdenes de los polinomios de las diferentes variables del modelo, a partir de:

$$y_t = v_1(B)x_{1t} + \cdots + v_k(B)x_{kt} + n_t$$

Una vez identificado el modelo, se procede a su estimación en forma similar al caso univariante, mediante un algoritmo de optimización para mínimos cuadrados no lineales.

Por su parte, la fase de diagnosis del modelo utilizará, asimismo, los desarrollos anteriormente expuestos para el caso univariante, a los que podemos añadir algunos propios del modelo de función de transferencia. En particular, una mala especificación del modelo puede traducirse en valores de x_t correlacionados con los de los residuos. Sin embargo, puede demostrarse que aun en el caso de que $\rho_{x,a}(k) = 0$, los coeficientes $r_{xa}(k)$ estimados presentarán una cierta correlación entre sí, lo que se traduce en que, con una función de autocorrelación cruzada que oscila alrededor de cero, sus variaciones se asemejan a las que corresponden a la función de autocorrelación de la propia x_t. Por ello, parece más recomendable contrastar la inadecuada especificación del modelo a partir de la función de autocorrelación cruzada entre a_t y α_t (es decir, entre los residuos del modelo de la función de transferencia y los correspondientes al modelo univariante de x_t) en lugar de entre x_t y a_t.

Etapas en la identificación, estimación y validación del modelo de función de transferencia

A continuación se resumen las etapas a seguir en los procesos de identificación, estimación y validación de un modelo de la función de transferencia.

1. ***Eliminación de tendencia en y_t y x_t.*** El orden d de la diferenciación será generalmente el mismo en ambas variables y sólo las series serán estacionarias, si decrecen rápidamente tanto las funciones de autocorrelación como la de correlación cruzada.

2. ***Análisis univariante de la serie x_t.*** Aplicación del proceso completo indicado previamente para el modelo estocástico univariante.

3. ***Preblanqueado de la serie y_t y estimación inicial de los coeficientes de la función de respuesta (v_k).***

4. ***Determinación de los órdenes de los polinomios de la función de transferencia, $w(B)$ y $\delta(B)$.*** A partir de la relación:

$$(1 - \partial_1 B - \cdots - \partial_r B^r)(v_0 + v_1 B + \cdots) = (w_0 - w_1 B - \cdots - w_s B^s)$$

puede tenerse una orientación sobre los posibles valores de r y s. Así, por ejemplo, para $r = 0$, el orden de s vendrá dado por el orden del mayor coeficiente v_1 no nulo. Concretamente, $s = 1 \Rightarrow v_0 + v_1 B + \cdots = w_0 - w_1 B$ con lo que sólo deberán ser estadísticamente distintos de cero los coeficientes estimados v_0 y v_1. Para el caso contrario de $s = 0$ y $r = 0$, como ejemplo complementario:

$$v_0 - v_1 B + \cdots = \frac{1}{1 - \partial B} = 1 + \partial B + \partial^2 \partial B^2 + \cdots$$

es decir, los coeficientes de la función de respuesta deberán seguir una progresión geométricamente decreciente. Un análisis detallado de diferentes valores de r y s para los casos más frecuentes de valores 0, 1 ó 2, nos evidenciaría las reglas de decisión a utilizar.

5. *Repetición del proceso para los distintos inputs x_{jt}, $j = 1, 2, \ldots, k$.*

6. *Determinación de los órdenes de los polinomios del término de error*, $\phi(B)$ y $\theta(B)$, bien a través del análisis univariante de y_t, bien estudiando el comportamiento de los residuos $n_t = y_t - v(B)x_t$.

7. *Estimación de los coeficientes del modelo* mediante el procedimiento de mínimos cuadrados no lineales con o sin la variante de predicción hacia atrás.

8. *Contraste de validez del modelo* a partir del estadístico t para la significación estadística de parámetros, matriz de correlación entre parámetros, contraste de ruido blanco y contraste de ausencia de correlación cruzada entre x_t y a_t o entre $\alpha_t\left[\alpha_t = \Psi^{-1}(B)x_t\right]$ y a_t.

Modelos de la función de transferencia estacionales

En el caso de que trabajemos con series estacionales, una expresión general del modelo de función de transferencia es la siguiente:

$$\nabla^{dY}\nabla^{DY}Y_t^{(\lambda_y)} = \sum_{j=1}^{k}\frac{w_j(B)\Omega_j(B^s)}{\delta_j(B)\Delta_j(B^s)}\nabla^{dX_j}\nabla^{DX_j}X^{(\lambda_{xj})} + \frac{\theta(B)\Theta(B^s)}{\phi(B)\Phi(B^s)}a_t$$

aunque, a efectos de aplicación práctica, es habitual que el anterior modelo se reduzca a la variante donde sólo el término de ruido es estacional, es decir:

$$\nabla^{dY}Y_t^{(\lambda_y)} = \sum_{j=1}^{k}\frac{w_j(B)}{\delta_j(B)}\nabla^{dX_j}X^{(\lambda_{xj})} + n_t$$

con:

$$\nabla^{dn}\nabla^{Dn}n_t = \frac{\theta(B)\Theta(B^s)}{\phi(B)\Phi(B^s)}a_t$$

Ejercicio 6-1. Consideramos las inversiones turísticas desde 1975 hasta 1986:

Años	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986
Inversión	600	800	750	400	350	500	1000	950	810	540	720	1160

Sabiendo que las inversiones turísticas son cíclicas con ciclo de 5 años, determinar la tendencia por el método de las medias móviles. Contestar a la misma pregunta considerando ciclo de 4 años.

Calcularemos medias móviles con $p = 5$ años. Como p es impar se forman medias relativas a los instantes $(p+1)/2$, $(p+3)/2$, $(p+5)/2$, ... (que serán valores enteros porque p es impar). La serie de medias móviles es la siguiente:

$$\bar{y}_{\frac{p+1}{2}} = \frac{y_1 + y_2 + \cdots y_p}{p},\quad \bar{y}_{\frac{p+3}{2}} = \frac{y_2 + y_3 + \cdots y_{p+1}}{p},\quad \bar{y}_{\frac{p+5}{2}} = \frac{y_3 + y_3 + \cdots y_{p+2}}{p},\ldots$$

A partir de los datos de nuestro problema tenemos:

$$\bar{y}_3 = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5},\quad \bar{y}_4 = \frac{y_2 + y_3 + y_4 + y_5 + y_6}{5},\ldots$$

La tabla siguiente muestra los cálculos.

Años	*Inversión*	*Media móvil*
1975	600	
1976	800	
1977	750	580
1978	400	560
1979	350	600
1980	500	640
1981	1000	722
1982	950	760
1983	810	804
1984	540	836
1985	720	
1986	1160	

La representación de la serie original y la ajustada por media móvil se observa en la Figura 6-24.

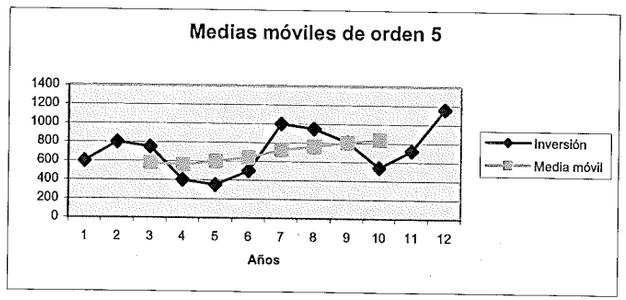


Figura 6-24

Consideramos ahora medias móviles con $p = 4$. Si p es par se forman medias relativas a los instantes $(p+1)/2$, $(p+3)/2$, $(p+5)/2$, ... (que no serán valores enteros porque p es par).

A continuación se hallan nuevas medias móviles entre cada dos medias móviles originales consecutivas, que serán ahora relativas a los instantes $(p+2)/2$, $(p+4)/2$, $(p+6)/2$, ... (que ya serán valores enteros porque p es par). La serie de medias móviles es la siguiente:

$$\bar{y}_{\frac{p+2}{2}} = \frac{\bar{y}_{\frac{p+1}{2}} + \bar{y}_{\frac{p+3}{2}}}{p}, \quad \bar{y}_{\frac{p+4}{2}} = \frac{\bar{y}_{\frac{p+3}{2}} + \bar{y}_{\frac{p+5}{2}}}{p}, \ldots$$

A partir de los datos de nuestro problema tenemos:

$$\bar{y}_{\frac{4+1}{2}} = \frac{y_1 + y_2 + y_3 + y_4}{4} = \bar{y}_{2,5}, \quad \bar{y}_{\frac{4+3}{2}} = \frac{y_2 + y_3 + y_4 + y_5}{5} = \bar{y}_{3,5}, \ldots$$

Como esta serie de medias móviles queda descentrada; será necesario centrarla obteniendo nuevas medias móviles de cada 2 de ellas. La tabla siguiente muestra los cálculos.

Años	*Inversión*	*Media móvil*	*Media móvil centrada*
1975	600		
1976	800	637,5	
1977	750	575	606,25
1978	400	500	537,5
1979	350	562,5	531,25
1980	500	700	631,25
1981	1000	815	757,5
1982	950	825	820
1983	810	755	790
1984	540	807,5	781,25
1985	720		
1986	1160		

Si representamos la serie inicial y las medias móviles de órdenes 4 y 5 sobre los mismos ejes, tenemos el gráfico de la Figura 6-25.

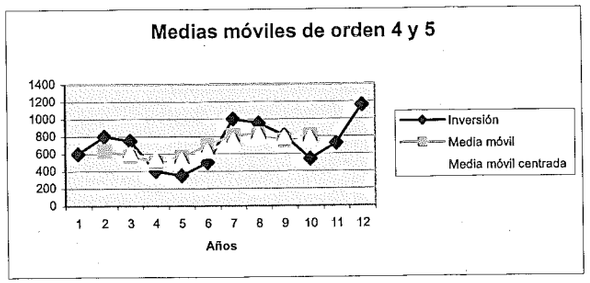


Figura 6-25

Ejercicio 6-2. Los datos siguientes reflejan la evolución de las ventas totales anuales de una empresa:

Años (t)	*Ventas* (Y_t)
1972	2,5
1973	3,1
1974	3,9
1975	4,7
1976	5,9
1977	7,0
1978	8,0
1979	9,6
1980	11,7
1981	14,0
1982	16,8
1983	19,2
	106,4

Con estos datos obtener las previsiones de ventas para 1985 suponiendo inicialmente crecimiento lineal y después crecimiento parabólico de segundo grado.

Se trata de realizar predicciones condicionales mediante modelos causales mediante un modelo de regresión lineal y otro parabólico. Comenzamos realizando el cambio de variable $t' = t - 1977$ para simplificar los cálculos y a continuación realizamos el ajuste lineal $y_t = a + bt'$ planteando el sistema de ecuaciones normales siguiente:

$$\left.\begin{array}{l} \sum_{t=1}^{T} y_t = Ta + b\sum_{t=1}^{T} t' \\ \sum_{t=1}^{T} t' y_t = a\sum_{t=1}^{T} t' + b\sum_{t=1}^{T} t'^2 \end{array}\right\} \Rightarrow \quad \left.\begin{array}{l} 106{,}4 = 12a + 6b \\ 265{,}6 = 6a + 146b \end{array}\right\} \Rightarrow a = 8{,}12 \text{ y } b = 1{,}49$$

La recta ajustada será $y_t = 8{,}12 + 1{,}49\, t' \Rightarrow y_t = 8{,}12 + 1{,}49\,(t\text{-}1977)$

La predicción lineal para 1985 será:

$$Y_{1985} = 8{,}12 + 1{,}49\,(1985\text{-}1977) = 20{,}4$$

A continuación realizamos el ajuste parabólico de segundo grado $y_t = a + bt' + c\,t'^2$ planteando el sistema de ecuaciones normales siguiente:

$$\left.\begin{aligned}\sum_{t=1}^{T} y_t &= Ta + b\sum_{t=1}^{T} t' + c\sum_{t=1}^{T} t'^2\\ \sum_{t=1}^{T} t'y_t &= a\sum_{t=1}^{T} t' + b\sum_{t=1}^{T} t'^2 + c\sum_{t=1}^{T} t'^3\\ \sum_{t=1}^{T} t'^2 y_t &= a\sum_{t=1}^{T} t'^2 + b\sum_{t=1}^{T} t'^3 + c\sum_{t=1}^{T} t'^4\end{aligned}\right\} \Rightarrow \left.\begin{aligned}106{,}4 &= 12a + 6b + 146c\\ 265{,}6 &= 6a + 146b + 216c\\ 1658{,}8 &= 146a + 216b + 3254c\end{aligned}\right\} \Rightarrow$$

$$a = 6{,}97,\ \ b = 1{,}37 \text{ y } c = 0{,}11$$

La parábola ajustada será $y_t = 6{,}97 + 1{,}37\, t' + 0{,}11\, t'^2$, o lo que es lo mismo:

$$y_t = 6{,}97 + 1{,}37(t\text{-}1977) + 0{,}11\,(t\text{-}1977)^2$$

La predicción parabólica para 1985 será:

$$Y_{1985} = 6{,}97 + 1{,}37(1985\text{-}1977) + 0{,}11\,(1985\text{-}1977)^2 = 24{,}97$$

Los cálculos usados en el problema se presentan en la tabla siguiente:

t	y_t	t'	t'^2	$t'y_t$	t'^3	t'^4	$t'^2 y_t$
1972	2,5	-5	25	-12,5	-125	625	62,5
1973	3,1	-4	16	-12,4	-64	256	49,6
1974	3,9	-3	9	-11,7	-27	81	35,1
1975	4,7	-2	4	-9,4	-8	16	18,8
1976	5,9	-1	1	-5,9	-1	1	5,9
1977	7	0	0	0,0	0	0	0
1978	8	1	1	8,0	1	1	8
1979	9,6	2	4	19,2	8	16	38,4
1980	11,7	3	9	35,1	27	81	105,3
1981	14	4	16	56,0	64	256	224
1982	16,8	5	25	84,0	125	625	420
1983	19,2	6	36	115,2	216	1296	691,2
	106,4	6	146	265,6	216	3254	1658,8

Ejercicio 6-3. Una región de un determinado país ha experimentado en siete años las siguientes entradas de turistas por estaciones en millones

	1978	*1979*	*1980*	*1981*	*1982*	*1983*	*1984*
Primavera	*2*	*3*	*3*	*4*	*4*	*5*	*6*
Verano	*6*	*9*	*11*	*14*	*16*	*19*	*20*
Otoño	*3*	*3*	*4*	*4*	*5*	*5*	*6*
Invierno	*1*	*1*	*2*	*2*	*3*	*3*	*4*

Analizar la estacionalidad de la serie y, si es necesario, desestacionalizarla a través de los índices de variación estacional (método de la tendencia). Representar simultáneamente la serie original y la serie desestacionalizada.

Comenzamos graficando la serie de datos originales para observar la estacionalidad (Figura 6-26).

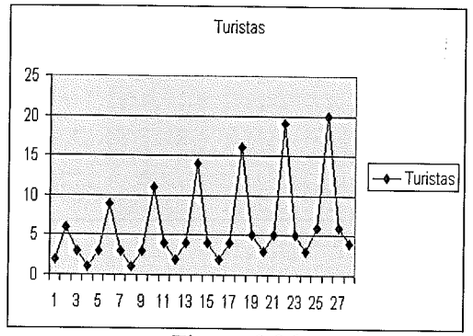


Figura 6-26

Para eliminar la estacionalidad, se comienza calculando la serie de medias anuales $\bar{y}_{i.}$, y ajustándola a una recta $\bar{y}_{i.} = a + bi$. Se obtiene la recta ajustada $\bar{y}_{i.} = 2 + i$ con pendiente $b = 1$. Observamos que $\bar{y}_{i.} = \{3,4,5,6,7,8,9\}$ cuya media es 6, y que $i = \{1,2,3,4,5,6,7\}$ cuya media es 4, con lo que la recta de ajuste tendrá de ecuación $\bar{y}_{i.} - 6 = (\sigma_{iy}/\sigma_i^2)(i\text{-}4)$. Resulta que tanto la covarianza de $\bar{y}_{i.}$ e i como la varianza de i valen 4, lo que nos lleva a la recta de ajuste $\bar{y}_{i.} - 6 = (4/4)(i\text{-}4) \Leftrightarrow \bar{y}_{i.} = 2 + i$.

Luego se calculan las *medias por estación* (trimestrales) $\bar{y}_{.k}$ $k = 1, 2, 3, 4$, y se aísla la componente estacional, obteniendo la serie de *medias trimestrales corregidas* $\bar{y}'_{.k} = \bar{y}_{.k} - b(k-1)/4$, cuya media $\bar{y}'$ sirve para hallar los *índices de variación estacional* $I_k = 100\bar{y}'_{.k} / \bar{y}'$ denotados por IVE. Por último, se obtiene la *serie desestacionaliza* dividiendo sus valores por los índices de variación estacional.

A continuación se presenta una tabla con toda la información.

C	D	E	F	G	H	I	J	K	L	M
	1978	1979	1980	1981	1982	1983	1984	Medias por estación	Medias corregidas	IVE
	2	3	3	4	4	5	6	3,857142857	3,857142857	68,5714286
	6	9	11	14	16	19	20	13,57142857	13,32142857	236,825397
	3	3	4	4	5	5	6	4,285714286	3,785714286	67,3015873
	1	1	2	2	3	3	4	2,285714286	1,535714286	27,3015873
Medias anuales=	3	4	5	6	7	8	9		5,625	
		Serie desestacionalizada								
	1978	1979	1980	1981	1982	1983	1984			
	2,92	4,38	4,38	5,83	5,83	7,29	8,75			
	2,53	3,8	4,64	5,91	6,76	8,02	8,45			
	4,46	4,46	5,94	5,94	7,43	7,43	8,92			
	3,66	3,66	7,33	7,33	11	11	14,7			

A continuación se representan la serie original y desestacionalizada en los mismos ejes (Figura 6-27).

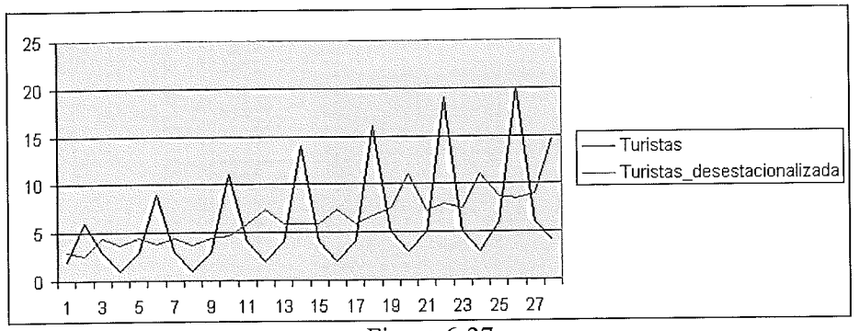


Figura 6-27

Ejercicio 6-4. Consideramos la producción de una empresa en 17 periodos cuyos datos son los siguientes:

Período (t)	*Producción* (X_t)
1	67
2	65
3	72
4	86
5	73
6	77
7	86
8	92
9	85
10	102
11	106
12	115
13	107
14	109
15	116
16	120
17	125

Utilizar el método de Holt con a = 0,2, b = 0,3 y b_1 = 3,4 para obtener predicciones de producción con un horizonte de tres periodos.

Según el método atopredictivo determinista de Holt, los valores predichos vienen dados por el modelo lineal definido como $F_t(l) = S_{t-1} + (b_{t-1})l \quad t > 2$, donde $S_t = aX_t + (1-a)[S_{t-1} + b_{t-1}]$ y $b_t = b[S_t - S_{t-1}] + (1-b)b_{t-1}$. Los valores iniciales son $S_1 = x_1$ y $b_1 = x_2-x_1$, pero el valor de b_1 puede ser otro ($b_1 = (x_3-x_1)/2$, $b_1 = (x_4-x_1)/3$, etc.). En nuestro problema nos dan b_1 = 3,4. Entonces, para nuestros datos tenemos las condiciones iniciales $S_1 = x_1 = 67$ y $b_1 = 3{,}4$.

Calcularemos S_2 y b_2 como sigue:

$$S_2 = aX_2 + (1-a)[S_1 + b_1] = 0{,}2(65) + (1-0{,}2)[67 + 3{,}4] = 69{,}32$$
$$b_2 = b[S_2 - S_1] + (1-b)b_1 = 0{,}3(69{,}32-65)+(1-0{,}3)3{,}4 = 3{,}076$$

Calcularemos S_3 y b_3 como sigue:

$$S_3 = aX_3 + (1-a)[S_2 + b_2] = 0{,}2(72) + (1-0{,}2)[69{,}32 + 3{,}076] = 72{,}3168$$
$$b_3 = b[S_3 - S_2] + (1-b)b_2 = 0{,}3(72{,}3168-69{,}32)+(1-0{,}3)3{,}076 = 3{,}05224$$

De la misma forma se calculan el resto de los valores S_t y b_t (t = 3,4,...,17).

El cálculo de las predicciones es inmediato mediante $F_t(l) = S_{t-1} + (b_{t-1})l \quad t > 2$.

La tabla siguiente resume los resultados.

t	X_t	S_t	b_t	*PREDICCIONES*	
1	67	67	3,4		
2	65	69,32	3,076		
3	72	72,3168	3,05224	72,396	l=1
4	86	77,495232	3,6900976	75,36904	l=1
5	73	79,5482637	3,19897782	81,1853296	l=1
6	77	81,5977932	2,85414333	82,7472415	l=1
7	86	84,7615492	2,94702714	84,45193654	l=1
8	92	88,5668611	3,20451256	87,70857637	l=1
9	85	90,4170989	2,79823014	91,77137366	l=1
10	102	94,9722633	3,3253104	93,21532906	l=1
11	106	99,8380589	3,78745598	98,29757365	l=1
12	115	105,900412	4,46992508	103,6255149	l=1
13	107	109,69627	4,26770486	110,370337	l=1
14	109	112,97118	3,9698664	113,9639745	l=1
15	116	116,752837	3,91340364	116,941046	l=1
16	120	120,532992	3,87342921	120,6662404	l=1
17	125	124,525137	3,90904392	124,4064215	l=1
18				128,4341812	l=1
19				132,3432251	l=2
20				136,252269	l=3

La predicción para $t = 18$ se calcula mediante:

$F_{18}(1) = S_{17} + (b_{17})(1) = 124{,}525137 + 3{,}90904392(1) = 128{,}4341812$

La predicción para $t = 19$ se calcula mediante:

$F_{19}(1) = S_{17} + (b_{17})(2) = 124{,}525137 + 3{,}90904392(2) = 132{,}3432251$

La predicción para $t = 19$ se calcula mediante:

$F_{20}(1) = S_{17} + (b_{17})(3) = 124{,}525137 + 3{,}90904392(3) = 136{,}252269$

Si representamos sobre los mismos ejes la serie original (que es ajustable fácilmente a una recta para justificar el uso de un método de predicción con modelo lineal como el de Holt) y la serie de predicciones tenemos la Figura 6-28.

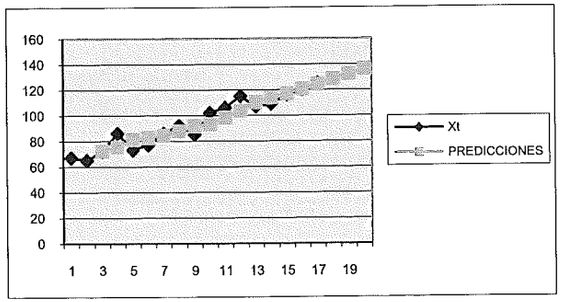


Figura 6-28

Ejercicio 6-5. Supongamos que la serie de ventas de una empresa evoluciona en el tiempo como sigue:

Periodo	1	2	3	4	5	6	7	8	9	10	11	12	13
Ventas	95	76	83	92	87	105	107	96	110	98	115	106	117

Utilizar el método de Brown de tendencia lineal con a = 0,1 para obtener predicciones de ventas con un horizonte de tres periodos.

Se tiene $S'_t(l) = aX_t + (1 - a)S'_{t-1}(l)$ y $S''_t(l) = aS'_t + (1 - a)S'_{t-1}(l)$ para el *modelo simple de alisado exponencial de Brown*. Se toma como valores iniciales $S'_1 = S''_1 = X_1 = 95$. Como valor fijado para a se toma 0,1.

La predicción en el periodo t a horizonte l se obtendrá mediante $S_t(l) = a_t + b_t l$, donde a_t y b_t son $a_t = 2S'_t - S''_t$ y $b_t = a(S'_t - S''_t)/(1-a)$.

Según las fórmulas anteriores, S'_2, S''_2, a_2 y b_2 se obtienen como sigue:

$S'_2(1) = aX_2 + (1 - a)S'_1(1) = 0{,}1(76) + 0{,}9(95) = 93{,}1$
$S''_2(1) = aS'_2 + (1 - a)S'_1(1) = 0{,}1(93{,}1) + 0{,}9(95) = 94{,}8$
$a_2 = 2S'_2 - S''_2 = 2(93{,}1) - 94{,}8 = 91{,}4$

$b_2 = a(S'_2 - S''_2)/(1-a) = 0{,}1(93\text{-}1\text{-}94{,}8) = -0{,}19$

La predicción en el periodo t a horizonte l es $S_t(l) = a_{t-1} + b_{t-1}l$, luego:

$S_3(1) = a_2 + b_2(1) = 91{,}39 + (-0{,}19) = 91{,}2$

El resto de los valores se calcula de forma similar obteniéndose la tabla siguiente:

t	X_t	S'_t	S''_t	a_t	b_t	S_t	
1	95	95	95				
2	76	93,1	94,81	91,39	-0,19		
3	83	92,09	94,538	89,642	-0,272	91,2	l=1
4	92	92,081	94,2923	89,8697	-0,2457	89,37	l=1
5	87	91,5729	94,0204	89,12544	-0,27194	89,624	l=1
6	105	92,9156	93,9099	91,92134	-0,11048	88,8535	l=1
7	107	94,324	93,9513	94,6968	0,041416	91,81086	l=1
8	96	94,4916	94,0053	94,97795	0,054034	94,73821	l=1
9	110	96,0425	94,2091	97,87591	0,203714	95,03199	l=1
10	98	96,2382	94,412	98,0645	0,202918	98,07962	l=1
11	115	98,1144	94,7822	101,4466	0,370244	98,26741	l=1
12	106	98,903	95,1943	102,6116	0,412076	101,8168	l=1
13	117	100,713	95,7461	105,6792	0,551838	103,0237	l=1
						106,2311	l=1
						106,7829	l=2
						107,3347	l=3

Las tres predicciones pedidas se han obtenido de la siguiente forma:

$S_{14} = a_{13} + b_{13}(1) = 105{,}6792 + 0{,}551838 = 106{,}2311$
$S_{15} = a_{13} + b_{13}(2) = 105{,}6792 + 0{,}551838(2) = 106{,}7829$
$S_{16} = a_{13} + b_{13}(3) = 105{,}6792 + 0{,}551838(3) = 107{,}3347$

Si representamos sobre los mismos ejes la serie original (ajustable a una recta para justificar el uso de un método de predicción con tendencia lineal como el de Brown) y la serie de predicciones tenemos el gráfico de la Figura 6-29.

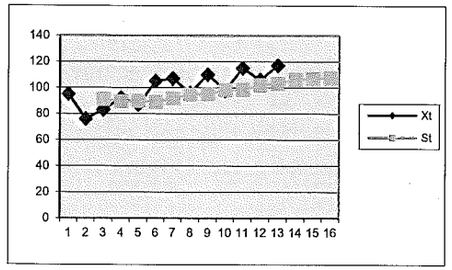


Figura 6-29

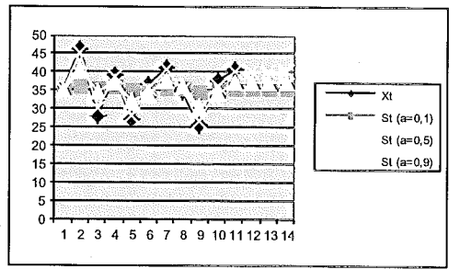


Figura 6-30

Ejercicicio 6-6. Consideramos una serie temporal con los siguientes datos:

t	1	2	3	4	5	6	7	8	9	10	11
X_t	35,2	46,1	28,3	39,4	26,8	36,7	41,2	34,6	25,4	37,4	40,5

Utilizar el modelo simple de alisado exponencial de Brown alternativamente con a = 0,1, a = 0,5 y a = 0,9 para obtener predicciones de ventas con un horizonte de tres periodos. Comparar las predicciones obtenidas para los distintos valores del parámetro a.

En general podemos poner $S_t(l) = aX_t + (1 - a)S_{t-1}(l)$ para el *modelo simple de alisado exponencial de Brown*. Se toma como valor inicial $S_1 = X_1$. Para $a = 0{,}1$ tenemos:

$$S_2(1) = aX_2 + (1 - a)S_1(1) = 0{,}1(46{,}1) + 0{,}9(35{,}2) = 36{,}3$$
$$S_3(1) = aX_3 + (1 - a)S_2(1) = 0{,}1(28{,}3) + 0{,}9(36{,}3) = 35{,}5$$

De la misma forma se obtiene el resto de los términos de la serie de predicciones, tanto para $a = 0{,}1$ como para $a = 0{,}5$ y $a = 0{,}9$. Se obtienen los resultados de la tabla siguiente:

t	X_t	S_t (a = 0,1)	S_t (a = 0,5)	S_t (a = 0,9)
1	35,2	35,2	35,2	35,2
2	46,1	36,29	40,65	45,01
3	28,3	35,491	34,475	29,971
4	39,4	35,8819	36,9375	38,4571
5	26,8	34,97371	31,86875	27,96571
6	36,7	35,146339	34,284375	35,826571
7	41,2	35,7517051	37,7421875	40,6626571
8	34,6	35,6365346	36,1710938	35,2062657
9	25,4	34,6128811	30,7855469	26,3806266
10	37,4	34,891593	34,0927734	36,2980627
11	40,5	35,4524337	37,2963867	40,0798063
12		35,4524337	37,2963867	40,0798063
13		35,4524337	37,2963867	40,0798063
14		35,4524337	37,2963867	40,0798063

Se observa que más allá del periodo 11 no existen más observaciones y las predicciones son las mismas para cualquier horizonte. Si representamos sobre el mismo gráfico las series inicial y de predicciones, se observa que el valor del parámetro a tiene una notable incidencia sobre las predicciones (Figura 6-30). Se observa que valores más pequeños de a alisan más los datos.

Ejercicicio 6-7. Se considera una serie de ventas con datos menuales desde enero de 1988 hsta abril de1996.Con la finalidad de obtener 6 predicciones se utiliza un modelo ARIMA cuyos resultados se presentan a continuación en este ejercicio obtenidos mediante el software SAS. Se trata de comentar las fases de identificación, estimación, diagnosis y predicción de este modelo.

La salida de SAS a comentar es la siguiente:

Proc ARIMA

Nombre de la variable = VENTAS

Periodo(s) de diferenciación	1
Media de series de trabajo	0.660589
Desviación estándar	2.011543
Número de observaciones	99
Observación(es) eliminadas por diferenciación	1

Comprobación de autocorrelación del ruido blanco

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	154.44	6	<.0001	0.828	0.591	0.454	0.369	0.281	0.198
12	173.66	12	<.0001	0.151	0.081	-0.039	-0.141	-0.210	-0.274
18	209.64	18	<.0001	-0.305	-0.271	-0.218	-0.183	-0.174	-0.161
24	218.04	24	<.0001	-0.144	-0.141	-0.125	-0.085	-0.040	-0.032

Estimación por mínimos cuadrados condicional

Parámetro	Estimador	Error estándar	Valor t	Approx Pr > |t|	Retardo
MU	0.89288	0.49391	1.81	0.0738	0
MA1,1	-0.58935	0.08988	-6.56	<.0001	1
AR1,1	0.74755	0.07785	9.60	<.0001	1

Constante Estimación	0.225409
Varianza Estimación	0.904034
Error std Estimación	0.950807
AIC	273.9155
SBC	281.7009
Número de residuales	99

* AIC y SBC no incluyen determinante de la log.

Correlaciones de los estimadores de parámetro

Parámetro	MU	MA1,1	AR1,1
MU	1.000	0.030	0.107
MA1,1	0.030	1.000	0.395
AR1,1	0.107	0.395	1.000

Comprobación de los residuales de autocorrelación

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	3.95	4	0.4127	0.016	-0.044	-0.068	0.145	0.024	-0.094
12	7.03	10	0.7227	0.088	0.087	-0.037	-0.075	0.051	-0.053
18	15.41	16	0.4951	-0.221	-0.033	-0.092	0.086	-0.074	-0.005
24	16.96	22	0.7657	0.011	-0.066	-0.022	-0.032	0.062	-0.047

```
                    Modelo para la variable VENTAS

Media estimada                          0.892875
Periodo(s) de diferenciación                   1


                    Factores autorregresivos

                Factor 1:  1 - 0.74755 B**(1)


                    Factores de la media móvil

                Factor 1:  1 + 0.58935 B**(1)

              Predicciones para la variable VENTAS

                          Error
Obs     Predicción         std.     95% Límites de confianza

101       171.0320        0.9508      169.1684        172.8955
102       174.7534        2.4168      170.0165        179.4903
103       177.7608        3.9879      169.9445        185.5770
104       180.2343        5.5658      169.3256        191.1430
105       182.3088        7.1033      168.3866        196.2310
106       184.0850        8.5789      167.2707        200.8993
```

En cuanto a la fase de identificación, lo primero que hace el software es mostrar que la serie ha sido diferenciada una vez y presentar estadísticos básicos. A continuación secomprueba que la serie no es un ruido blanco, porque de lo contrario no será modelizable. Los p-valores del contraste chi-cuadrado de ruido blanco son muy pequeños, lo que nos permite aceptar la hipótesis de que la serie no es un conjunto de números aleatorios y es posible realizar predicciones.

Por otro lado, al observar los resultados de la estimación del modelo, vemos que SAS ha identificado un modelo ARIMA(1,1,1) con constante ya que ofrece las estimaciones de la constante MU y los términos MA1,1 y AR1,1 del modelo ARIMA.

En cuanto a la fase de estimación, dado que la serie inicial se identifica como un modelo ARIMA(1,1,1) y viendo los valores de los parámetros estimados, podemos escribir la ecuación algebraica del modelo estimado como:

$$(1\text{-}0{,}74755B)DVENTAS = 0{,}89288 + (1+0{,}58935B)\ RESID \Rightarrow$$

$$(1\text{-}0{,}74755B)(1\text{-}B)VENTAS = 0{,}89288 + (1+0{,}58935B)\ RESID \Rightarrow$$

$$(1\text{-}1{,}74755B+0{,}74755B^2)VENTAS = 0{,}89288 +(1+0{,}58935B)\ RESID \Rightarrow$$

$$X_t - 1{,}74755X_{t-1} +0{,}74755X_{t-2} = 0{,}89288 +e_t+0{,}58935e_{t-1} \qquad (X_t = \text{VENTAS})$$

Ya tenemos entonces la forma algebraica definitiva del modelo ARIMA(1,1,1) ajustado para la serie inicial. Esta expresión permitirá el cálculo de las predicciones conocidos los dos primeros valores de la serie y el primer valor residual.

En cuanto a la diagnosis, se observa que los parámetros estimados del modelo son muy significativos (la constante tiene un p-valor de 0,0738 y los términos AR y MA tienen un p-valor menor que 0,0001).

Por otro lado, la matriz de correlaciones de los parámetros estimados tiene valores prqueños, lo cual siempre es deseable.

Por otra parte, el contraste chi-cuadrado de autocorrelación de los residuos presenta p-valores altos, lo que indica que los residuos son un ruido blanco, condición esencial para validar un modelo ARIMA.

También se observan las predicciones relativas a los 6 meses del año siguiente con sus intervalos de confianza al 95%. Estas predicciones son de calidad debido a la buena diagnosis del modelo identificado y estimado. Se observa que los intervalos de confianza de las predicciones son más anchos a medida que avanzamos en el horizonte de predicción.

Por último, es necesario tener también presente que la serie tiene la longitud suficiente para aplicar la metodoligía ARIMA.

Ejercicicio 6-8. Se considera una serie X_t de ratios de productividad empresarial con datos menuales desde enero de 1951 hasta febrero de1996.Con la finalidad de obtener 10 predicciones se utiliza un modelo ARIMA cuyos resultados se presentan a continuación en este ejercicio obtenidos mediante el software SPSS. Se trata de comentar las fases de identificación, estimación, diagnosis y predicción de este modelo.

La salida de SPSS es la siguiente:

Model Statistics

Model	Number of Predictors	Model Fit statistics			Ljung-Box Q(18)			Number of Outliers
		R-squared	RMSE	Normalized BIC	Statistics	DF	Sig.	
ratio-Modelo_1	0	,651	,074	-5,141	22,303	16	,100	0

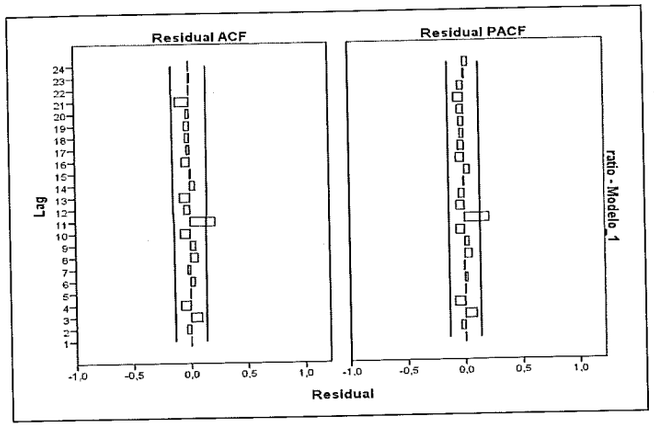
ARIMA Model Parameters

					Estimate	SE	t	Sig.
ratio-Modelo_1	ratio	Natural Log	AR	Lag 1	,905	,043	20,924	,000
			MA	Lag 1	,531	,084	6,302	,000
			Seasonal Difference		1			
			MA, Seasonal	Lag 1	,632	,069	9,111	,000

Forecast

Model		Mar 1969	Apr 1969	May 1969	Jun 1969	Jul 1969	Aug 1969	Sep 1969	Oct 1969	Nov 1969	Dec 1969
ratio-Modelo_1	Forecast	1,22	1,11	1,13	,99	1,13	1,12	1,47	1,26	1,21	1,31
	UCL	1,37	1,25	1,28	1,13	1,29	1,29	1,68	1,45	1,40	1,51
	LCL	1,09	,98	,99	,87	,98	,98	1,27	1,09	1,05	1,13

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.



Al observar los resultados de la estimación del modelo en la tabla ARIMA *Model Parameters*, vemos que SPSS ha identificado un modelo ARIMA(1,0,1)(0,1,1)$_{12}$ sin constante para la serie en logaritmos naturales. Lógicamente se han aplicado logaritmos por los problemas de estacionariadad en varianza de la serie original y también fue necesario diferenciar una vez la parte estacional para estacinarizar finalmente la serie inicial.

En cuanto a la fase de estimación, dado que la serie inicial en logatimos LX se identifica como un modelo ARIMA(1,0,1)(0,1,1)$_{12}$ y viendo los valores de los parámetros estimados, podemos escribir la ecuación algebraica del modelo estimado como:

$$(1\text{-}0{,}89B)DLX = (1+0{,}51B)(1+0{,}68B^{12})RESID$$

$$DLSAS\text{-}0{,}89DLX(-1) = RESID+0{,}51RESID(-1)+0{,}68RESID(-12) +0{,}51*0{,}68RESID(-13)$$

$$DLX = 0{,}89DLX(-1) + RESID+0{,}51RESID(-1)+0{,}68RESID(-12) +0{,}51*0{,}68RESID(-13)$$

No olvidemos que $DLX = Ln(X_t)\text{-}Ln(X_{t-12})$, con lo que podemos escribir (LX=Ln(X_t)):

$$Ln(X_t)\text{-}Ln(X_{t-12}) = 0{,}89[Ln(X_t)\text{-}Ln(X_{t-12})\text{-}(Ln(X_{t-1})\text{-}Ln(X_{t-13}))]+ e_t+0{,}51e_{t-1}+0{,}68e_{t-12}+0{,}346e_{t-13}$$

$$0{,}11\ Ln(X_t)+0{,}89\ Ln(X_{t-1})\text{-}0{,}11\ Ln(X_{t-12})\text{-}0{,}89\ Ln(X_{t-13}) = e_t+0{,}51e_{t-1}+0{,}68e_{t-12}+0{,}346e_{t-13}$$

Ya tenemos entonces la forma algebraica definitiva del modelo ARIMA ajustado para la serie inicial. Esta expresión permitirá el cálculo de las predicciones conocidos los dos primeros valores de la serie y el primer valor residual.

En cuanto a la diagnosis, los resultados de la estimación de la tabla ARIMA *Model Parameters* son correctos, ya que los parámetros estimados tienen una significatividad muy alta (p-valores muy pequeños). En la tabla *Model Statistics* también se observan otros estadísticos de ajuste como un R^2 de valor 0,651 y un p-valor del estadístico de Ljung-Box mayor que 0,05 que indica aleatoriedad de los residuos al 95%, así como la ausencia de valores atipicos. Las gráficas de las funciones de autocorrelación residual y autocorrelación parcial residual corroboran este hecho ya que prácticamente todos sus términos están dentro de las franjas de confianza al 95%.

En la tabla *Forecast* se observan las 10 predicciones con sus correspondientes intervalos de confianza. La calidad de las predicciones es alta porque la fase de diagnosis del modelo se ha superado muy satisfactoriamente.

Ejercicicio 6-9. Se considera una serie Z_t de pasajeros de líneas aéreas con datos mensuales desde enero de 1949 hasta diciembre de1960. Con la finalidad de obtener predicciones de calidad se utiliza un modelo ARIMA con análisis de la intervención cuyos resultados se presentan a continuación en este ejercicio obtenidos mediante el software Tramo/Seats a través de Eviews. Se trata de comentar las fases de identificación, estimación, diagnosis de este modelo.

El resumen de la salida que presenta Tramo/Seats a través de Eviews se comenta paso a paso como se indica a continuación:

```
AUTOMATIC MODEL IDENTIFICATION BEGINS

MODEL FINALLY CHOSEN:

  (0,1,1)(0,1,1)

 WITHOUT MEAN

 WITHOUT TRADING DAY CORRECTION

 WITHOUT EASTER CORRECTION
```

Esta primera parte de la salida constituye el proceso de identificación de la parte ARIMA del modelo. Se observa que el modelo identificado automáticamente tiene una estructura ARIMA(0,1,1)(0,1,1)$_{12}$ sin constante. No se identifica corrección de efecto fin de semana ni de efecto semana santa. A continuación comienza el proceso de estimación.

```
ARIMA MODEL ESTIMATION BEGINS

   INITIAL PARAMETER VALUES:
          -.29843
          -.46497
    ITERATION, LAMBDA    1    0.0000
  F0 FP    0.12457        0.12427
  F0-FP SUM S   0.30073E-03   0.29606E-03    1.0158

  ITERATION, LAMBDA    2    0.0000
```

BIBLIOTECA CALI

```
METHOD OF ESTIMATION: EXACT MAXIMUM LIKELIHOOD
PARAMETER        ESTIMATE          STD ERROR        T RATIO    LAG

MA1  1           -.33105           0.85248E-01      -3.88       1
MA2  1           -.49591           0.86042E-01      -5.76      12

REGULAR MA INVERSE ROOTS ARE
NO.      REAL P.      IMAG.P.      MODULUS     ARGUMENT      PERIOD
 1      0.33105       0.0000      0.33105      0.0000           -
SEASONAL MA INVERSE ROOTS ARE
NO.      REAL P.      IMAG.P.      MODULUS     ARGUMENT      PERIOD
 1      0.49591       0.0000      0.49591      0.0000           -

CORRELATIONS OF THE ESTIMATES

 1.000   -0.142
-0.142    1.000

AIC
   -526.073

BIC
    -6.7268

FINAL VALUE OF OBJECTIVE FUNCTION:
             0.12426

ITERATIONS:            2

NUMBER OF FUNCTION EVALUATIONS:          7
```

Se han obtenido las estimaciones de los parámetros del modelo ARIMA adecuado a la serie, su desviación típica y el contraste individual de la T con valores adecuados (los suficientemente pequeños). Además se incluyen las raíces de los polinomios estimados para la parte regular y la parte estacional y la matriz de correlaciones entre los parámetros obtenidos. Se incluyen los criterios informativos AIC y BIC. Todos los valores indican un buen proceso de estimación. A continuación comienza el proceso de análisis de la intervención mediante la identificación de las diferentes tipologías de valores atípicos.

```
ESTIMATES OF REGRESSION PARAMETERS
CONCENTRATED OUT OF THE LIKELIHOOD

PARAMETER            VALUE          ST. ERROR      T VALUE
MU                  -.14943E-03    (     0.00102)   -0.15
OUT 1 (135)         -.10324        (     0.02548)   -4.05   AO    ( 3 1960)
OUT 2 ( 29)          0.95931E-01   (     0.02227)    4.31   AO    ( 5 1951)
OUT 3 ( 62)         -.80285E-01    (     0.02204)   -3.64   AO    ( 2 1954)
OUT 4 ( 54)         -.96760E-01    (     0.02558)   -3.78   LS    ( 6 1953)

COVARIANCE MATRIX OF ESTIMATORS

 0.104E-05  -0.866E-10  -0.611E-08   0.148E-07   0.165E-06
-0.866E-10   0.649E-03  -0.818E-07  -0.246E-05  -0.245E-06
-0.611E-08  -0.818E-07   0.496E-03  -0.961E-06  -0.428E-04
 0.148E-07  -0.246E-05  -0.961E-06   0.486E-03  -0.294E-05
 0.165E-06  -0.245E-06  -0.428E-04  -0.294E-05   0.654E-03
```


El proceso de identificación, contraste y estimación de valores atípicos (*outliers*) muestra la presencia de tres outliers de tipo aditivo AO (en marzo de 1960, en mayo de 1951 y en febrero de 1954) y un cambio de nivel LS (en junio de 1953). Según la información anterior, el modelo ARIMA con intervención estimado tendrá la ecuación de predicción siguientesiguiente:

$$\boxed{\begin{gathered}(1-B)(1-B^{12})\log(z_t) = (1-0.33105B)(1-0{,}49591B^{12})a_t \\ -0.00014943 - 0.10324I_t^{1960} - 0.095931I_t^{1951} - 0.080285I_t^{1954} - \frac{0{,}09676}{1-B}I_t^{1953}\end{gathered}}$$

$$I_t^{t_0} = 1 \text{ si } t = t_0 \text{ y } I_t^{t_0} = 0 \text{ si } t \neq t_0$$

A continuación comienza el prceso de diagnosis residual

```
TEST-STATISTICS ON RESIDUALS
 ----------------------------

          MEAN=  -0.0022517
       ST.DEV.=   0.0027532
       OF MEAN
       T-VALUE= -0.8178
 NORMALITY TEST=   1.541        ( CHI-SQUARED(2) )

      SKEWNESS=  0.2113          ( SE =  0.2182 )
      KURTOSIS=  2.6609          ( SE =  0.4364 )

 SUM OF SQUARES=   0.1209815
  DURBIN-WATSON=  1.9477
 STANDARD ERROR=   0.3123552E-01
 OF RESID.
  MSE OF RESID.=   0.9756575E-03

  LJUNG-BOX Q VALUE OF ORDER 24 IS    21.61 AND IF RESIDUALS ARE RANDOM IT
SHOULD BE DISTRIBUTED AS CHI-SQUARED(22)

APPROXIMATE TEST OF RUNS ON RESIDUALS
 -------------------------------------
  NUM.DATA= 126
   NUM.(+)=  63
   NUM.(-)=  63
  NUM.RUNS=  58
   T-VALUE= -1.0733

 APPROXIMATE TEST OF RUNS ON AUTOCORRELATION FUNCTION
 ----------------------------------------------------
  NUM.DATA=  36
   NUM.(+)=  18
   NUM.(-)=  18
  NUM.RUNS=  18
   T-VALUE= -0.3382
```

Se observa que los estadísticos de normalidad residual presentan valores correctos (valor es de los coeficientes de asimetría y curtosis prácticamente en el intervalo [-2 ,2]). No hay problemas de autocorrelación residual porque el estadístico de Durbin Watson vale 1,9477 que está muy próximo al valor ideal 2.

El estadístico Q de LJUNG-BOX (calculado para 24 retardos) vale 21,61 para una chi-cuadrado de 22 grados de libertad. Como el valor crítico al 95% para una chi-cuadrado de 22 grados de libertad es 33,924, aceptamos la hipótesis nula de ausencia de correlaciones relevantes al 95% de confianza, con lo que podemos suponer que los residuos siguen un ruido blanco. También se realiza el test de rachas residual que tiene valores de la T de Student adecuados (lo suficientemente pequeños).

CAPÍTULO 7

HERRAMIENTAS PARA EL ANÁLISIS UNIVARIANTE DE SERIES TEMPORALES.

EVIEWS Y LA IDENTIFICACIÓN, ESTIMACIÓN, VALIDACIÓN Y PREDICCIÓN DE MODELOS ARIMA(*p,d,q*)(*P,D,Q*)*s*

Eviews permite llevar a cabo el proceso completo de la metodología Box y Jenkins. Para ello utilizamos el archivo *arimas.wf*1 que contiene una variable de nombre *SA* que representa una serie de ratios mensuales sobre la producción de una empresa. Con la finalidad de realizar predicciones de producción futuras se trata de ajustar la serie de ratios a un modelo ARIMA general convenientemente.

Para comenzar la fase de identificación, y con el objeto de observar la estacionalidad, realizamos una representación gráfica de la serie mediante *Quick* → *Graph* → *Line Graph* (Figura 7-1), indicando la serie a graficar en *Series List* (Figura 7-2) para obtener la representación de la serie en la Figura 7-3. Se observa a simple vista que el gráfico presenta estructura estacional mensual. Sin embargo, este hecho hay que comprobarlo de modo formal.

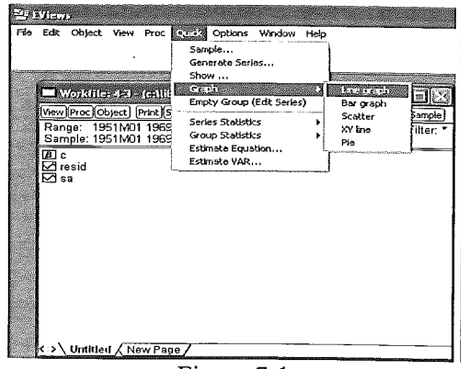


Figura 7-1

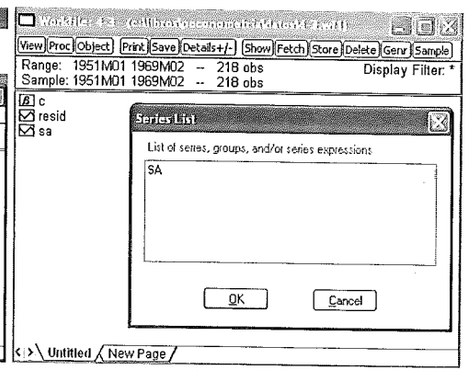


Figura 7-2

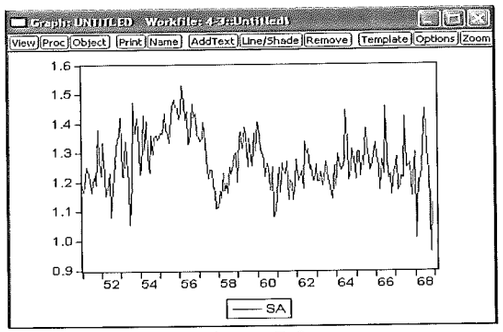


Figura 7-3

Para probar la estacionalidad podemos utilizar el ***gráfico vertical de la serie*** (Figura 7-5), que se obtiene haciendo doble clic sobre la serie *SA* para ver sus valores y eligiendo *View* → *Graph* → *Spike* (Figura 7-4), el ***gráfico de las subseries estacionales*** (Figura 7-81) que se obtiene eligiendo *View* → *Graph* → *Seasonal Stacked Line* (Figura 7-6) o el ***gráfico de las subseries anuales*** (Figura 7-7) que se obtiene eligiendo *View* → *Graph* → *Seasonal Split Line* (Figura 7-8).

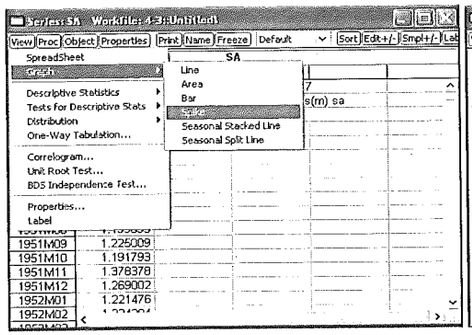


Figura 7-4

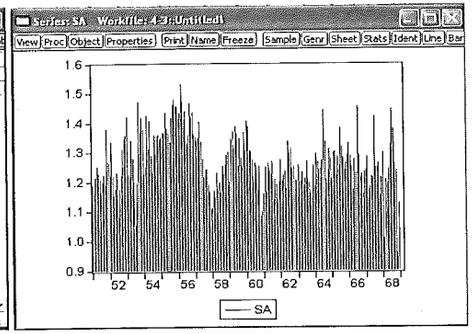


Figura 7-5

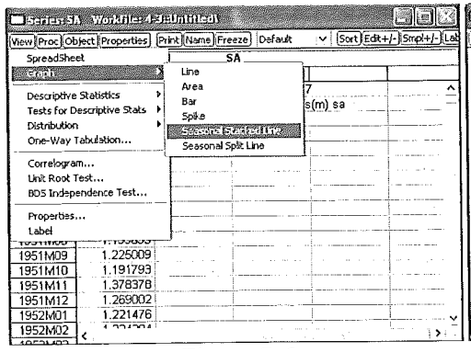


Figura 7-6

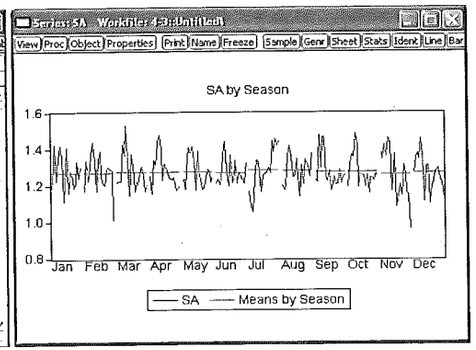


Figura 7-7

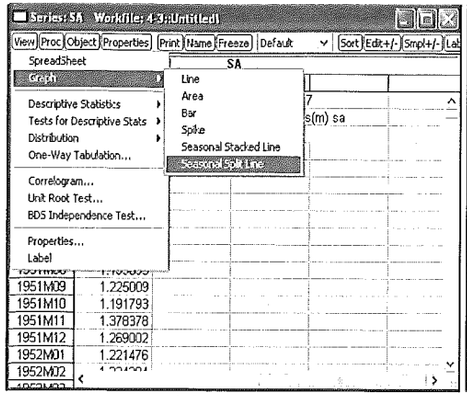


Figura 7-8

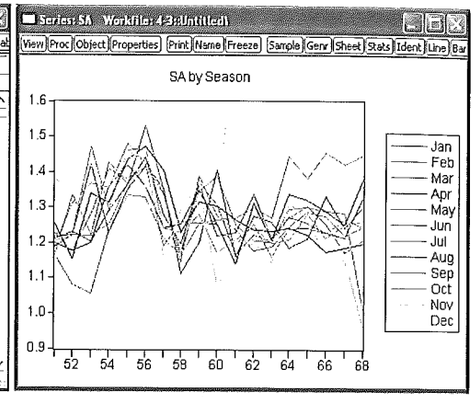


Figura 7-9

Todos los gráficos anteriores muestran claramente la presencia de estacionalidad mensual. El gráfico de las subseries anuales presenta evoluciones paralelas de los datos en los distintos meses de todo los años. El gráfico de las subseries estacionales muestra claramente las secciones similares de las estaciones.

Pero la estacionalidad, así como la estacionariedad, también puede detectarse a través de las funciones de autocorrelación y autocorrelación parcial estimadas (FAC y FACP respectivamente). Para ello elegimos *View* → *Correlogram* (Figura 7-10) y elegimos la serie en niveles con 36 retardos (Figura 7-11). Se obtienen las funciones de autocorrelación y autocorrelación parcial estimadas de la Figura 7-12. La FAC muestra valores altos en los retardos múltiplos del periodo estacional 12, 24 y 36.

Se observa que las funciones de autocorrelación y autocorrelación parcial estimadas también validan los periodos estacionales porque los coeficientes de la FAC para retardos múltiplos del periodo estacional de la serie son significativamente distintos de cero.

Además, para una cantidad grande de retardos la FAC se configura en forma de abanico que completa su ciclo girando sobre el eje de abscisas para una cantidad de retardos igual al periodo estacional. Por otro lado, la FACP presenta estructura de coeficientes significativos para retardos periódicos (largos). La FAC y la FACP deben considerarse a la vez, pues a veces intercambian sus papeles en el comportamiento estacional. Asimismo, los coeficientes de la FAC no decaen rápidamente, lo que indica falta de estacionariedad en media. Asimismo, si consideramos la serie de medias y varianzas por meses a lo largo de toda la serie observamos variaciones significativas crecientes y decrecientes a lo largo de los años, lo que indica que no hay estacionariedad ni en media ni en varianza en la serie original.

Descriptivos

	AÑO		Estadístico	Error típ.
SA	1951	Media	1,2277	,02893
		Varianza	,010	
	1952	Media	1,2353	,03822
		Varianza	,018	
	1953	Media	1,2930	,04233
		Varianza	,022	
	1954	Media	1,3319	,03140
		Varianza	,012	
	1955	Media	1,3914	,02761
		Varianza	,009	
	1956	Media	1,4346	,02574
		Varianza	,008	
	1957	Media	1,3131	,02794
		Varianza	,009	
	1958	Media	1,1852	,02338
		Varianza	,007	
	1959	Media	1,3099	,01918
		Varianza	,004	
	1960	Media	1,2959	,03378
		Varianza	,014	
	1961	Media	1,1922	,02226
		Varianza	,006	
	1962	Media	1,2436	,02666
		Varianza	,009	
	1963	Media	1,2145	,02385
		Varianza	,007	
	1964	Media	1,2643	,03400
		Varianza	,014	
	1965	Media	1,2864	,03118
		Varianza	,012	
	1966	Media	1,2791	,03899
		Varianza	,018	
	1967	Media	1,2454	,03687
		Varianza	,016	
	1968	Media	1,2515	,04210
		Varianza	,021	
	1969	Media	1,0997	,08457
		Varianza	,014	

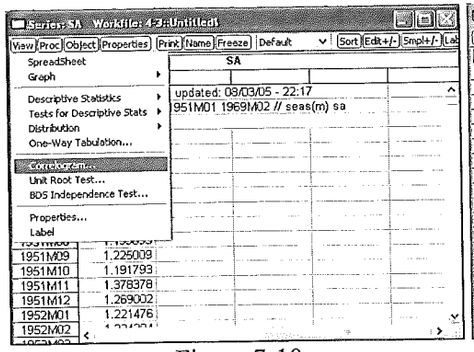


Figura 7-10

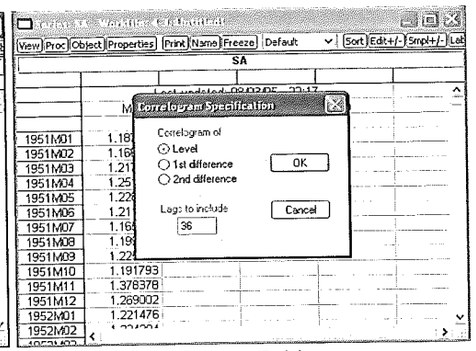


Figura 7-11

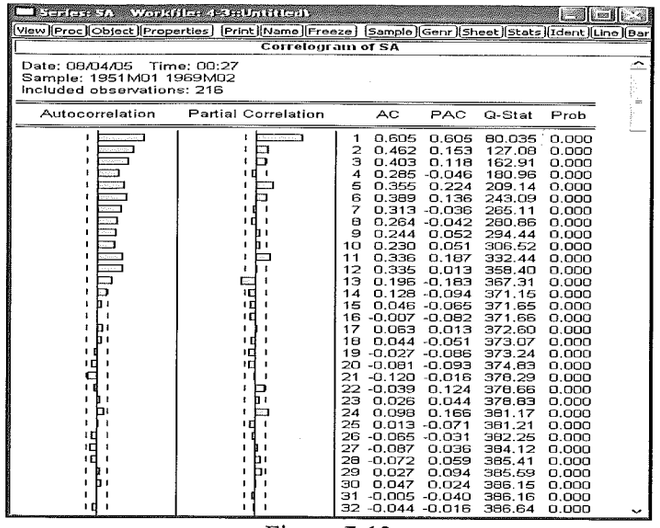


Figura 7-12

Este resultado nos lleva a tomar logaritmos de la serie original (botón GENR y Figura 7-13) y diferenciar la serie de logaritmos con el objetivo de atenuar la falta de estacionariedad en media y en varianza. Pero, una vez aplicados los logaritmos, como la serie es estacional, el problema es identificar si diferenciamos la parte regular de la serie en logaritmos (botón GENR y Figura 7-14) o en la parte estacional (botón GENR y Figura 7-15). Para ello representamos las funciones de autocorrelación estimada y autocorrelación parcial estimada bajo los supuestos de diferenciación en la parte regular de la serie en logaritmos (Figuras 7-16 y 7-17) o en la parte estacional (Figura 7-18).

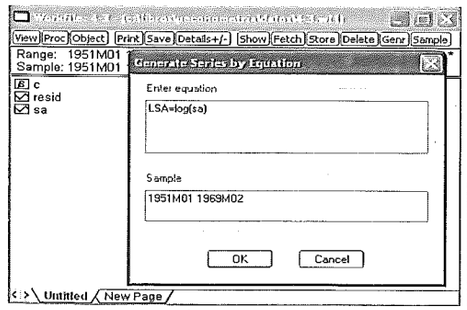


Figura 7-13

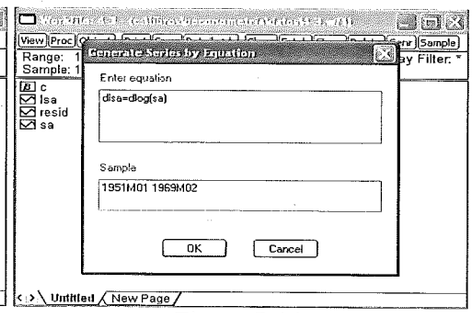


Figura 7-14

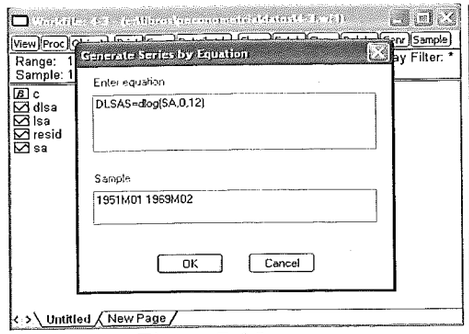


Figura 7-15

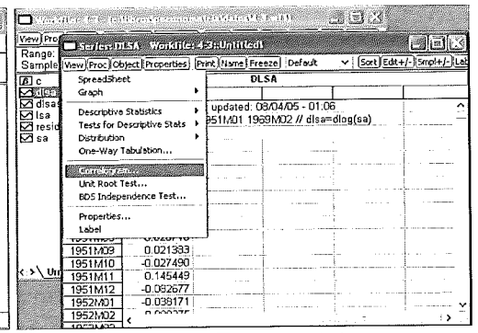


Figura 7-16

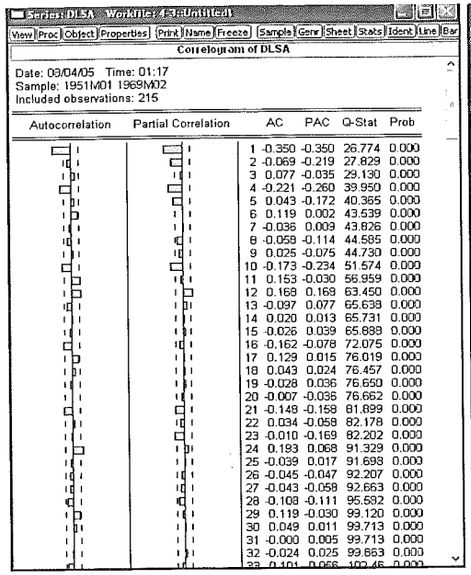


Figura 7-17

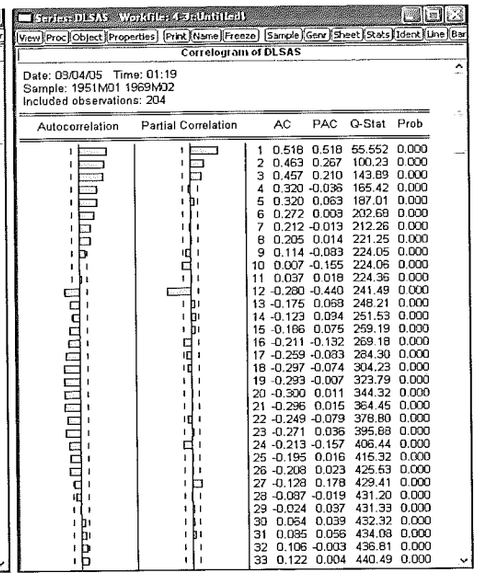


Figura 7-18

Se observa que al diferenciar sólo la parte regular de la serie en logaritmos, las funciones de autocorrelación y autocorrelación parcial estimadas (Figura 7-17) no superan el problema de la falta de estacionariedad ya que la FAC no decae rápidamente. Pero al diferenciar sólo una vez la parte estacional de la serie en logaritmos, las funciones de autocorrelación y autocorrelación parcial estimadas (Figura 7-18) ya superan el problema de la no estacionariedad. Asimismo, estas dos funciones cumplen las condiciones para que haya estacionalidad porque los coeficientes de la FAC para retardos múltiplos del periodo estacional de la serie son significativamente distintos de cero. Además, para una cantidad grande de retardos, la FAC se configura en forma de abanico que completa su ciclo girando sobre el eje de abscisas para una cantidad de retardos igual al periodo estacional.

Luego el problema de la estacionalidad y la estacionariedad en media y en varianza se ha arreglado aplicando logaritmos, diferenciando una vez la parte estacional y no diferenciando la parte regular. Luego la parte regular de la serie en logaritmos es integrada de orden cero $I(0)$ y la parte estacional es integrada de orden uno $I(1)$.

Ahora resta por identificar el orden de la parte autorregresiva AR y la parte de medias móviles MA. Para realizar esta tarea utilizaremos la FAC y la FACP para las que se obtuvo la estacionariedad y la estacionalidad, es decir las representadas en la Figura 7-92. Observando estas dos funciones vemos que sus coeficientes no se anulan bruscamente con periodicidades y que sus estructuras se ajustan claramente a un modelo $ARMA(1,1)(0,1)_{12}$. Existe una amplia literatura con baterías de gráficos de FAC y FACP parcial ajustadas a sus correspondientes modelos ARMA. En nuestro caso, la parte AR(1) de la parte regular proviene del decrecimiento rápido inicial y las ondas sinusoidales de la FAC añadido a que la FACP presenta sólo un coeficiente significativo en la mayoría de los periodos (salvo en el primero), anulándose bruscamente el resto de los coeficientes. Asimismo, la parte MA(1) de la parte regular proviene de que la FAC presenta un solo retardo significativo en la mayoría de los periodos (salvo en el primero). De todas formas, la única duda posible sería considerar también AR(1) la parte estacional.

Ya tenemos identificada completamente la serie inicial como un modelo $ARIMA(1,0,1)(0,1,1)_{12}$. Es decir, ya hemos realizado el trabajo más importante en la modelización de una serie temporal mediante la metodología de Box-Jenkins.

Una vez identificado el modelo realizamos su estimación y diagnosis. Para ello se elige *Quick → Estimate Equation*, se escribe la ecuación del modelo a ajustar en el campo *Equation Specification* de la solapa *Specification* teniendo en cuenta la estructura ARIMA previamente identificada, se elige *LS - Least Squares (NLS and ARIMA)* en el campo *Method* (Figura 7-19) y se hace clic en *Aceptar*. Se obtienen los resultados de la Figura 7-20. El modelo presenta buena significatividad individual y conjunta de los parámetros estimados, altos coeficientes de determinación y un estadístico de Durbin Watson casi igual a 2. Luego la diagnosis del ajuste es correcta.

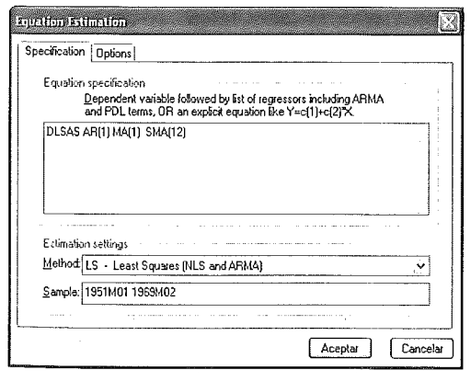


Figura 7-19

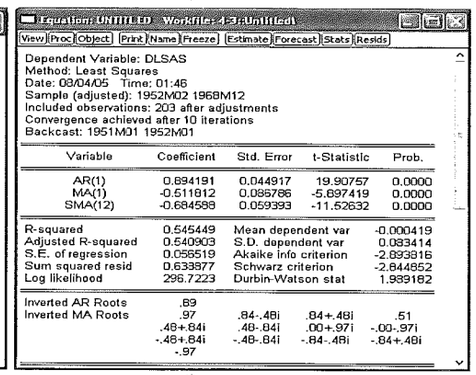


Figura 7-20

También es un buen instrumento de diagnosis el correlograma residual (Figura 7-22) obtenido mediante *View → Residual Tests → Correlogram Q-Statistics* (Figura 7-21). Se observa que tanto la FAC como la FACP no tienen retardos claramente significativos y además ***las probabilidades asociadas al estadístico Q son casi todas mayores que 0,05***, lo que indica que ***los residuos del modelo estimado se comportan como un ruido blanco.***

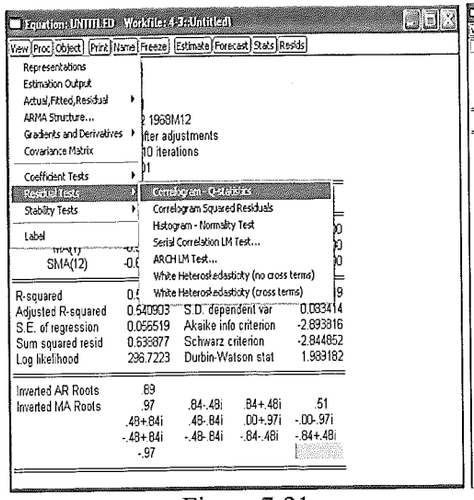


Figura 7-21

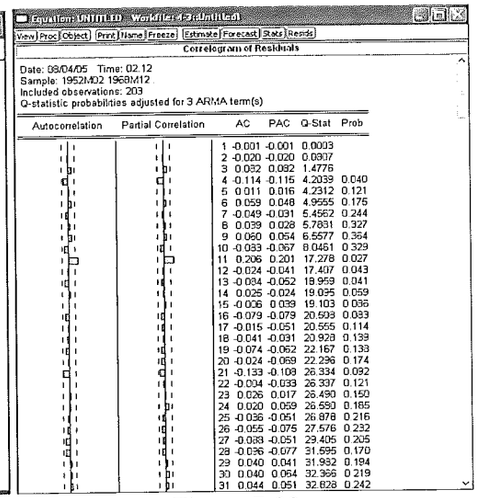


Figura 7-22

Dado que la serie inicial como un modelo ARIMA(1,0,1)(0,1,1)$_{12}$ podemos escribir la ecuación algebraica del modelo como:

$$(7-0{,}89B)DLSAS = (1+0{,}51B)(1+0{,}68B^{12})RESID$$

$$DLSAS-0{,}89DLSAS(-1) = RESID+0{,}51RESID(-1)+0{,}68RESID(-12) +0{,}51*0{,}68RESID(-13)$$

$$DLSAS = 0{,}89DLSAS(-1) + RESID+0{,}51RESID(-1)+0{,}68RESID(-12) +0{,}51*0{,}68RESID(-13)$$

No olvidemos que $DLSAS = Ln(X_t)-Ln(X_{t-12})$, con lo que podemos escribir ($SA=X_t$):

$$Ln(X_t)-Ln(X_{t-12}) = 0{,}89[Ln(X_t)-Ln(X_{t-12})-(Ln(X_{t-1})-Ln(X_{t-13}))]+ e_t+0{,}51e_{t-1}+0{,}68e_{t-12}+0{,}346e_{t-13}$$

$$0{,}11\ Ln(X_t)+0{,}89\ Ln(X_{t-1})-0{,}11\ Ln(X_{t-12})-0{,}89\ Ln(X_{t-13}) = e_t+0{,}51e_{t-1}+0{,}68e_{t-12}+0{,}346e_{t-13}$$

EVIEWS Y LOS MODELOS ARIMA Y DE INTERVENCIÓN

Tramo/Seats

Los programs TRAMO (*Time Series Regression with ARIMA Noise Missing Observations, and Outliers*) y SEATS (*Signal Extraction in ARIMA Time Series*) permiten ajustar modelos ARIMA automáticamente incluyendo análisis de la intervención. Dichos programas han sido incorporados a partir de la versión 5 de Eviews.

Como ejemplo ajustamos la serie temporal mensual x contenida en el archivo *estacional.wf1* con Tramo/Seats. Comenzamos cargando en memoria la serie mediante *Open → Eviews Workfile* (Figura 7-23). A continuación se hace doble clic sobre la serie que contiene los datos (x) y con la serie en pantalla se elige *Proc → Seasonal Adjustment → Tramo Seats* (Figura 7-24). Obtenemos la pantalla de opciones de la Figura 7-25. En la solapa *Outliers* elegimos detección automática *Autedetect all types* (Figura 7-26).

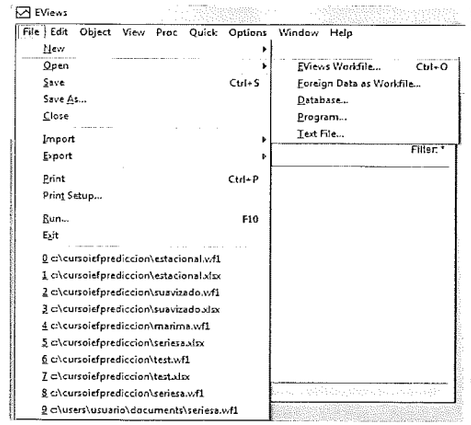


Figura 7-23

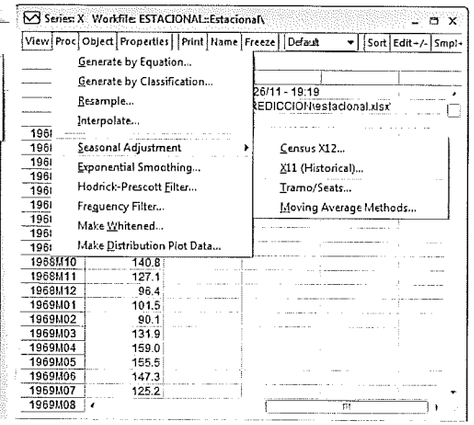


Figura 7-24

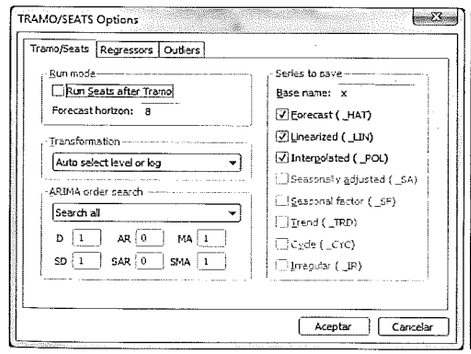


Figura 7-25

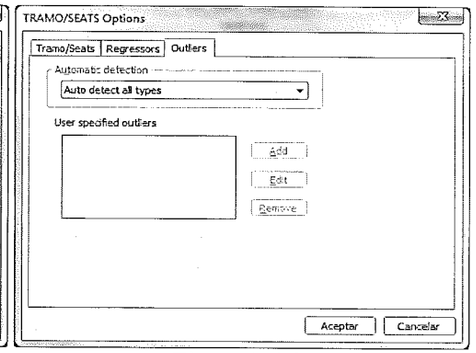


Figura 7-26

Al pulsar en *Aceptar* se obtiene la salida del procedimiento. La primera parte de la salida muestra la características del proceso de identificación y estimación del modelo ARIMA adecuado para la serie.

```
TIME SERIES REGRESSION MODELS WITH ARIMA ERRORS, MISSING VALUES AND OUTLIERS.
                          BETA VERSION (*)

                                BY

                   VICTOR GOMEZ & AGUSTIN MARAVALL

              with the programming assistance of G. CAPORELLO

(*) Copyright : V. GOMEZ, A. MARAVALL (1994,1996)

 SERIES TITLE=evtramo

 ORIGINAL SERIES

 NUMBER OF OBSERVATIONS: 166

 YEAR              JAN       FEB       MAR       APR       MAY       JUN
JUL       AUG       SEP       OCT       NOV       DEC

 1968            80.500    84.600   126.600   162.000   140.900   137.900
139.800   136.600   134.300   140.800   127.100    96.400
 1969           101.500    90.100   131.900   159.000   155.500   147.300
125.200   124.900   129.300   123.400    94.600    84.100
 1970            66.400    74.300   114.700   128.400   125.000   135.200
140.800   128.700   130.900   140.900   126.900   121.400
 1971           110.600   102.200   167.900   201.100   198.500   193.800
194.300   204.500   173.800   179.700   173.700   152.100
 1972           149.100   152.200   203.900   211.600   225.800   223.100
206.500   228.600   203.000   216.500   185.700   150.500
 1973           146.600   138.000   200.000   205.000   234.000   202.600
202.600   197.200   148.400   147.100   133.300    90.400
 1974            84.500   109.400   124.800   159.800   149.000   147.600
126.600   111.100    98.300    96.700    75.100    55.100
 1975            56.100    54.700    80.200    97.900   116.100   110.300
119.300   117.300   111.900   123.600    96.900    76.100
 1976            72.500    89.900   118.400   137.200   147.900   154.200
136.600   145.900   151.800   148.400   127.100   107.400
 1977            81.300   112.500   173.600   182.200   201.300   197.600
189.800   194.000   177.700   193.100   154.800   129.200
 1978            88.600   101.300   172.100   197.500   211.000   216.000
192.200   190.900   180.500   192.100   158.600   119.500
 1979            88.200    84.500   152.900   161.000   189.100   191.800
164.200   170.300   163.700   169.000   118.700    91.600
 1980            73.100    79.900    85.100    96.200    91.700   116.400
120.100   129.900   138.300   152.700   112.900    95.900
 1981            84.500    71.900   107.800   123.000   109.900   105.800
99.900    86.300    84.600    86.200
```

A continuación se presentan los parámetros especificados para la ejecución del programa.

```
MODEL PARAMETERS
----------------
 MQ= 12          IMEAN=  1       LAM= -1       D=  1       BD=  1
 P=  0           BP=  0          Q=  1         BQ=  1      IREG=  0
 ITRAD=  0       IEAST=  0       IDUR=  0      M= 36       QM= 24
```



```
 INCON=  0       NBACK=  0       NPRED=  8     INTERP=  2   INIT=  0
 IFILT=  2       IDENSC=  1      IROOT=  2     INIC=  3     ICONCE=  1
 ICDET=  1       IATIP=  1       IMVX=  0      IDIF=  3     PG=  0
 AIO=  0         INT1=  1        INT2= 166     RSA=  0      SEATS=  0
 VA= 3.50        TOL=  0.100E-03               PC=  0.143E+00
 NOADMISS= 1     BIAS= 1         SMTR= 0
 THTR= -0.400    RMOD=  0.500    MAXBIAS=  0.500

TH =        -0.10

BTH =        -0.10

NUMBER OF INITIAL OBS. =  13

 MEAN IS NOT SIGNIFICANT:
 IMEAN CHANGED TO 0

TRANSFORMED SERIES (LOGARITHMS OF THE DATA)
 YEAR              JAN       FEB       MAR       APR       MAY       JUN
JUL       AUG       SEP       OCT       NOV       DEC

 1968            4.388     4.438     4.841     5.088     4.948     4.927
4.940     4.917     4.900     4.947     4.845     4.569
 1969            4.620     4.501     4.882     5.069     5.047     4.992
4.830     4.828     4.862     4.815     4.550     4.432
 1970            4.196     4.308     4.742     4.855     4.828     4.907
4.947     4.857     4.874     4.948     4.843     4.799
 1971            4.706     4.627     5.123     5.304     5.291     5.267
5.269     5.321     5.158     5.191     5.157     5.025
 1972            5.005     5.025     5.318     5.355     5.420     5.408
5.330     5.432     5.313     5.378     5.224     5.014
 1973            4.988     4.927     5.298     5.323     5.455     5.311
5.311     5.284     5.000     4.991     4.893     4.504
 1974            4.437     4.695     4.827     5.074     5.004     4.995
4.841     4.710     4.588     4.572     4.319     4.009
 1975            4.027     4.002     4.385     4.584     4.754     4.703
4.782     4.765     4.718     4.817     4.574     4.332
 1976            4.284     4.499     4.774     4.921     4.997     5.038
4.917     4.983     5.023     5.000     4.845     4.677
 1977            4.398     4.723     5.157     5.205     5.305     5.286
5.246     5.268     5.180     5.263     5.042     4.861
 1978            4.484     4.618     5.148     5.286     5.352     5.375
5.259     5.252     5.196     5.258     5.066     4.783
 1979            4.480     4.437     5.030     5.081     5.242     5.256
5.101     5.138     5.098     5.130     4.777     4.517
 1980            4.292     4.381     4.444     4.566     4.519     4.757
4.788     4.867     4.929     5.028     4.727     4.563
 1981            4.437     4.275     4.680     4.812     4.700     4.662
4.604     4.458     4.438     4.457
```

A continuación se presenta la identificación del modelo como un ARIMA(0,1,1)(0,1,1) sin constante, sin efecto semana santa y sin efecto fin de semana. También se identifica un outlier de cambio temporal tipo TC en el mes de marzo de 1980. Se observan también las estimaciones de los parámetros del modelo de intervención adecuado a la serie, su desviación típica y el contraste individual de la *t*. Además, se incluyen las raíces de los polinomios estimados para la parte regular y la parte estacional y la matriz de correlaciones entre los parámetros obtenidos. Se incluyen los criterios informativos AIC y BIC.

```
AUTOMATIC MODEL IDENTIFICATION BEGINS

MODEL FINALLY CHOSEN:

 (0,1,1)(0,1,1)

WITHOUT MEAN

WITHOUT TRADING DAY CORRECTION

WITHOUT EASTER CORRECTION

OUTLIERS

147 TC     ( 3 1980)

METHOD OF ESTIMATION: EXACT MAXIMUM LIKELIHOOD

PARAMETER        ESTIMATE          STD ERROR        T RATIO    LAG

MA1  1           -.22810           0.79677E-01      -2.86       1
MA2  1           -.68414           0.75306E-01      -9.08      12


REGULAR MA INVERSE ROOTS ARE
NO.      REAL P.      IMAG.P.      MODULUS      ARGUMENT      PERIOD
 1      0.22810      0.0000       0.22810       0.0000          -

SEASONAL MA INVERSE ROOTS ARE
NO.      REAL P.      IMAG.P.      MODULUS      ARGUMENT      PERIOD
 1      0.68414      0.0000       0.68414       0.0000          -

CORRELATIONS OF THE ESTIMATES

 1.0000     0.1787
 0.1787     1.0000

AIC
  -252.5709

BIC
    -4.4791
  FINAL VALUE OF OBJECTIVE FUNCTION:
              1.6529

ITERATIONS:              3

 NUMBER OF FUNCTION EVALUATIONS:            10


ESTIMATES OF REGRESSION PARAMETERS
CONCENTRATED OUT OF THE LIKELIHOOD

PARAMETER              VALUE          ST. ERROR       T VALUE
OUT 1 (147)            -.38712        (    0.09131)   -4.24    TC     ( 3 1980)

COVARIANCE MATRIX OF ESTIMATORS

0.834E-02
```

La ecuación del modelo estimado es la siguiente:

$$\boxed{(1-B)(1-B^{12})x_t = (1-0.2281B)(1-0{,}68414B^{12})a_t - \frac{0{,}38712}{1-\delta B} I_t^{1980} \quad 0<\delta<1}$$

$$I_t^{t_0} = 1 \text{ si } t = t_0 \text{ y } I_t^{t_0} = 0 \text{ si } t \neq t_0$$

A continuación se presenta información sobre contrastes de ruido blanco para los residuos y otros contrastes estadísticos residuales para la diagnosis. Los estadísticos de normalidad residual presentan valores correctos. No hay problemas de autocorrelación residual porque el estadístico de Durbin Watson está muy próximo a 2. El estadístico Q de LJUNG-BOX (calculado para 24 retardos) vale 21,61 para una chi-cuadrado de 22 grados de libertad. Como el valor crítico al 95% para una chi-cuadrado de 22 grados de libertad es 33,924, aceptamos la hipótesis nula de ausencia de correlaciones relevantes al 95% de confianza, con lo que podemos suponer que los residuos siguen un ruido blanco. También se realiza el test de rachas residual.

```
 NUMBER OF WHITE NOISE RESIDUALS          152

 WHITE NOISE RESIDUALS

     -0.0483     -0.0603      0.0831     -0.0080     -0.1473     -0.0165
0.0388     -0.0687
     -0.1505      0.0967     -0.2012      0.1012      0.0629     -0.0807
0.0283      0.1160
      0.1393     -0.0408     -0.0034      0.0706      0.0954      0.1592
0.0486     -0.0866
      0.0655      0.0217      0.0473     -0.0208      0.0283      0.0984     -
0.1483     -0.0286
      0.1112      0.0246      0.0852      0.0535     -0.1333     -0.1674
0.0665      0.0063
     -0.0558      0.0974     -0.0464      0.0234     -0.0354     -0.0849
0.0265     -0.0515
     -0.0312     -0.1114      0.1118     -0.1117      0.0130     -0.0510     -
0.2222     -0.1006
      0.0043     -0.2277     -0.0620      0.2647     -0.1915      0.1075     -
0.0842      0.0220
     -0.1219     -0.1677     -0.0196     -0.0459     -0.1458     -0.1099
0.0531     -0.0811
      0.0605      0.0691      0.1819      0.0278      0.1515      0.0528
0.0999      0.1103
     -0.0577      0.0026     -0.0125      0.1738     -0.0136     -0.0174
0.0141      0.0868
     -0.0806      0.0768      0.1642     -0.0248      0.0262      0.0899     -
0.2183      0.1806
      0.1632     -0.0715      0.0206      0.0016      0.0126      0.0240     -
0.0216      0.0582
     -0.0311      0.0379     -0.2534     -0.0911      0.1588      0.0512
0.0034      0.0406
     -0.0588     -0.0277      0.0069      0.0239      0.0046     -0.0705     -
0.1217     -0.2273
      0.1340     -0.0453      0.0787      0.0360     -0.0770      0.0159
0.0292     -0.0085
     -0.1642     -0.0625     -0.0084     -0.0066      0.1039     -0.1097     -
0.2796      0.0982
      0.0985      0.0507      0.1056      0.0630     -0.0579      0.0569
0.1099     -0.2329
     -0.1503      0.0341     -0.1258     -0.1208     -0.0126     -0.1722     -
0.0322     -0.0434
```

```
TEST-STATISTICS ON RESIDUALS

        MEAN=  -0.0073423
        ST.DEV.=0.0082286
        OF MEAN
        T-VALUE= -0.8923

 NORMALITY TEST=   2.525          ( CHI-SQUARED(2) )

       SKEWNESS= -0.3119            ( SE =  0.1987 )
       KURTOSIS=  2.9026            ( SE =  0.3974 )

 SUM OF SQUARES=     1.572557

 DURBIN-WATSON=  2.0231

 STANDARD ERROR=    0.1023900
 OF RESID.
  MSE OF RESID.=    0.1048371E-01

 AUTOCORRELATIONS
 ----------------
     -0.0129  -0.0384   0.2147   0.0058   0.0038   0.1635   0.0159  -0.1645
0.1226  -0.0044  -0.0853   0.0745
 SE   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811
0.0811   0.0811   0.0811   0.0811
 Q      0.03     0.26     7.50     7.50     7.50    11.79    11.83    16.23
 18.69    18.69    19.90    20.83
 PV    -1.00    -1.00     0.01     0.02     0.06     0.02     0.04     0.01
  0.01     0.02     0.02     0.02
      0.1483  -0.0836   0.1292  -0.0489   0.0333   0.0459  -0.0800   0.0103  -
0.0792  -0.0270  -0.1166   0.0012
 SE   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811
0.0811   0.0811   0.0811   0.0811
 Q     24.54    25.72    28.57    28.98    29.18    29.54    30.67    30.69
 31.81    31.94    34.41    34.41
 PV     0.01     0.01     0.01     0.01     0.02     0.02     0.02     0.03
  0.03     0.04     0.03     0.04
     -0.0644  -0.0443   0.0581  -0.0564  -0.0934  -0.0128  -0.0198  -0.0546  -
0.1178   0.0089  -0.1084  -0.1332
 SE   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811
0.0811   0.0811   0.0811   0.0811
 Q     35.17    35.54    36.17    36.77    38.43    38.46    38.54    39.12
 41.85    41.86    44.21    47.79
 PV     0.05     0.06     0.07     0.08     0.07     0.09     0.11     0.12
  0.09     0.11     0.09     0.06

  LJUNG-BOX Q VALUE OF ORDER 24 IS    34.41 AND IF RESIDUALS ARE RANDOM IT
SHOULD BE DISTRIBUTED AS CHI-SQUARED(22)

 PARTIAL AUTOCORRELATIONS
 ------------------------
     -0.0129  -0.0386   0.2140   0.0087   0.0208   0.1242   0.0171  -0.1688
0.0692  -0.0248  -0.0253   0.0235
 SE   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811
0.0811   0.0811   0.0811   0.0811
      0.1704  -0.0279   0.1164  -0.1498   0.1181  -0.0624  -0.0955   0.0026  -
0.0676  -0.0335  -0.0895   0.0013
 SE   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811
0.0811   0.0811   0.0811   0.0811
     -0.0141  -0.0269   0.0636  -0.0378  -0.0521  -0.0987   0.0137  -0.0267  -
0.1628   0.0899  -0.1086  -0.0247
 SE   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811
0.0811   0.0811   0.0811   0.0811
```



```
 APPROXIMATE TEST OF RUNS ON RESIDUALS
 -------------------------------------
  NUM.DATA= 152
   NUM.(+)=  76
   NUM.(-)=  76
  NUM.RUNS=  79
   T-VALUE=  0.3255

 APPROXIMATE TEST OF RUNS ON AUTOCORRELATION FUNCTION
 ----------------------------------------------------
  NUM.DATA=  36
   NUM.(+)=  18
   NUM.(-)=  18
  NUM.RUNS=  22
   T-VALUE=  1.0146

 SQUARED RESIDUALS:
 ------------------

 AUTOCORRELATIONS
 ----------------

      0.0854   0.0451  -0.0520  -0.1022   0.0217   0.0354  -0.1522  -0.1107
0.0461  -0.0237   0.0535   0.0233
 SE   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811
0.0811   0.0811   0.0811   0.0811
 Q      1.13     1.45     1.87     3.53     3.60     3.80     7.54     9.53
  9.88     9.97    10.45    10.54
 PV    -1.00    -1.00     0.17     0.17     0.31     0.43     0.18     0.15
  0.20     0.27     0.32     0.39
      0.0142   0.0298   0.1077  -0.0469  -0.0604  -0.1002  -0.0712  -0.0297  -
0.1099   0.0029   0.0151   0.0704
 SE   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811
0.0811   0.0811   0.0811   0.0811
 Q     10.57    10.72    12.70    13.08    13.72    15.47    16.36    16.52
 18.68    18.68    18.72    19.63
 PV     0.48     0.55     0.47     0.52     0.55     0.49     0.50     0.56
  0.48     0.54     0.60     0.61
      0.0369  -0.0069  -0.0507   0.0648  -0.0039  -0.0011  -0.1141  -0.0913  -
0.0749  -0.0563  -0.0015   0.0322
 SE   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811   0.0811
0.0811   0.0811   0.0811   0.0811
 Q     19.88    19.89    20.37    21.16    21.16    21.16    23.69    25.31
 26.41    27.04    27.04    27.25
 PV     0.65     0.70     0.73     0.73     0.78     0.82     0.74     0.71
  0.70     0.72     0.76     0.79

  LJUNG-BOX Q VALUE OF ORDER 24 IS    19.63 AND IF RESIDUALS ARE RANDOM IT
SHOULD BE DISTRIBUTED AS CHI-SQUARED(22)
```

Por último se presentan las predicciones

```
 FORECASTS:

 ORIGIN:          166 NUMBER:          8

 OBS       FORECAST        STD ERROR          ACTUAL            RESIDUAL
       FORECAST          STD ERROR
       (TR. SERIES)
(
 ORIGINAL SERIES)
```

```
 167        4.20296          0.102396
67.2359            6.90280
 168        3.98379          0.129356
     54.1718              7.03684
 169        3.78533          0.151594
     44.5595              6.79394
 170        3.79639          0.170962
     45.1960              7.78363
 171        4.23895          0.189149
     70.5868              13.4717
 172        4.32997          0.204570
     77.5475              16.0313
 173        4.31290          0.219126
     76.4710              16.9600
 174        4.34078          0.232871
     78.8770              18.6200
```

EVIEWS Y LOS MÉTODOS AUTOPROYECTIVOS DETERMINISTAS: ALISADOS EXPONENCIALES Y DE HOLT-WINTERS

Eviews permite trabajar con los métodos autoproyectivos deterministas especificados en el capítulo anterior. Se pueden utilizar los métodos de alisado exponencial, doble alisado, alisado de Holt Winters aditivo y multiplicativo y alisado estacional.

Para ejecutar estos métodos comenzamos cargando en memoria la serie mediante *Open* → *Eviews Workfile*. A continuación se hace doble clic sobre la serie que contiene los datos (*x*) del archivo *tramo1.wf1* y con la serie en pantalla se elige *Proc* → *Exponenial Smoothing* (Figura 7-27). Obtenemos la pantalla de opciones de la Figura 7-28 en la que se puede elegir el método de alisado (simple, doble, no estacional de Holt Winters y estacional aditivo o multiplicativo de Holt Winters) modelo aditivo o multiplicativo, los parámetros de alisado, la longitud de la serie y el ciclo estacional.

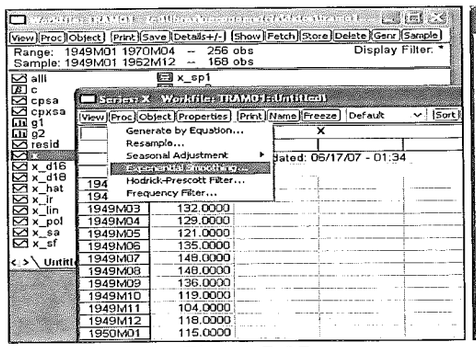


Figura 7-27

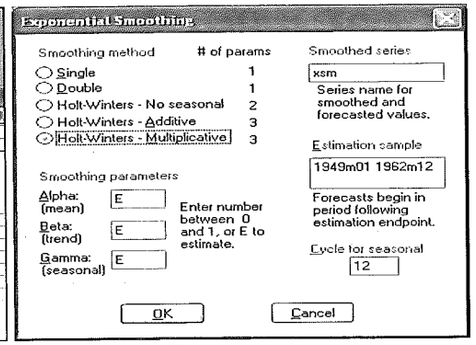


Figura 7-28

Al pulsar *Aceptar* obtenemos los valores predichos de un ciclo completo (12 meses) de la serie alisada por el método de Holt Winters estacional multiplicativo en nuestro caso (Figura 7-29).

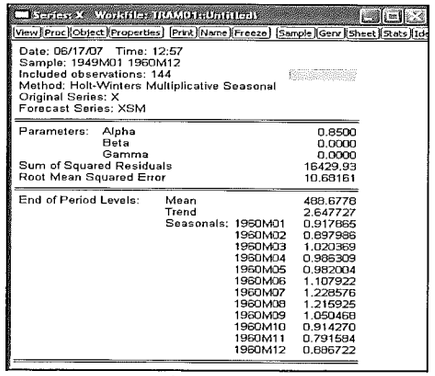


Figura 7-29

SPSS Y LA IDENTIFICACIÓN, ESTIMACIÓN, DIAGNOSIS Y PREDICCIÓN DE MODELOS ARIMA(*p,d,q*)(*P,D,Q*)*s*

SPSS dispone del procedimiento ARIMA que permite trabajar y realizar predicciones con series temporales a partir de la metodología de Box-Jenkins. Pero para llevar a cabo las tareas de identificación se utilizan especialmente las opciones de *Analizar* → *Predicciones* (Figura 7-30). Como ejemplo, ajustaremos a un modelo ARIMA adecuado una serie de ratios mensuales de solvencia de una empresa contenida en el archivo *arima.sav*. Para comenzar la fase de identificación realizamos una representación gráfica de la misma (*Analizar* → *Predicciones* → *Gráfico de secuencia*) con el objeto de observar la estacionalidad. Se rellena la pantalla de entrada de *Gráficos de secuencia* según la Figura 7-31 y se obtiene el gráfico de la Figura 7-32 que presenta estructura estacional mensual.

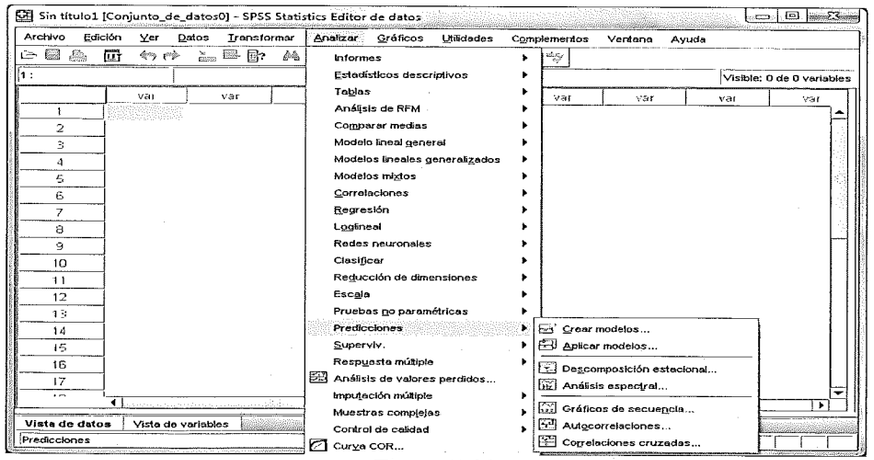


Figura 7-30

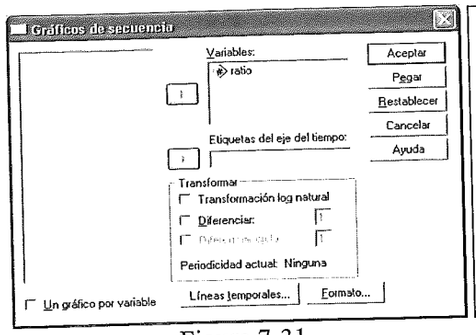


Figura 7-31

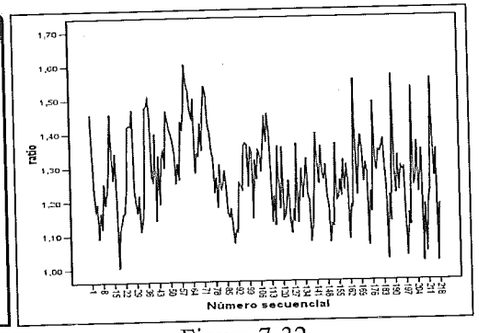


Figura 7-32

Ahora definimos la serie como estacional mensual, a partir de *Datos* → *Definir fechas*, rellenando la pantalla de entrada *Definir fechas* como se indica en la Figura 7-33. Al pulsar *Aceptar* se genera la serie mensual tal y como se observa en el *Editor de datos* (Figura 7-34).

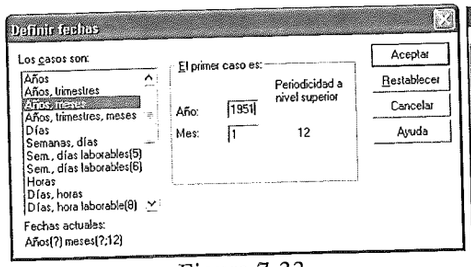


Figura 7-33

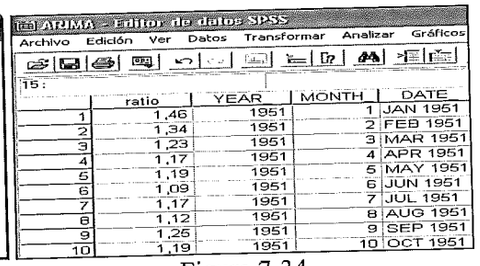


Figura 7-34

Para observar mejor la estacionalidad, representamos el periodograma por frecuencia de la serie mediante *Analizar* → *Predicciones* → *Análisis espectral* (Figura 7-30) y rellenando la pantalla de entrada del procedimiento *Diagramas espectrales* según se indica en la Figura 7-35. Al pulsar *Aceptar* se obtiene el periodograma de la Figura 7-36, cuyo segundo pico corresponde a la frecuencia 0,08, es decir la estación es 1/0,08 =12 meses.

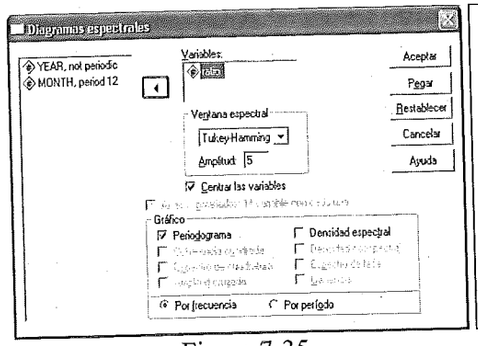


Figura 7-35

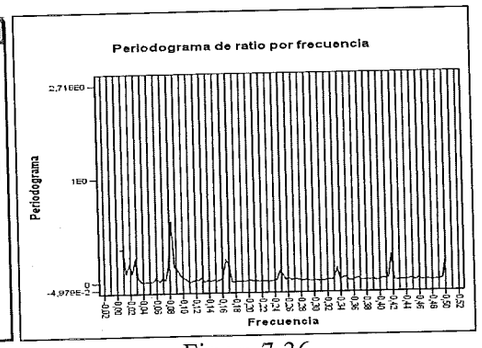


Figura 7-36

Pero la estacionalidad, así como la estacionariedad también pueden detectarse a través de las funciones de autocorrelación y autocorrelación parcial estimadas (FAC y FAC parcial respectivamente). Para ello elegimos *Analizar* → *Predicciones* → *Autocorrelaciones* (Figura 7-30) y rellenamos la pantalla de entrada del procedimiento *Autocorrelaciones* tal y como se indica en la Figura 7-37. En el botón *Opciones* situamos el número máximo de retardos a 36 para representar la FAC con un tramo significativo (Figura 7-38). Al pulsar *Aceptar* se obtiene la FAC de la Figura 7-39 que muestra valores altos en los retardos 12, 24 y 36. La función de autocorrelación parcial se presenta en la Figura 7-40.

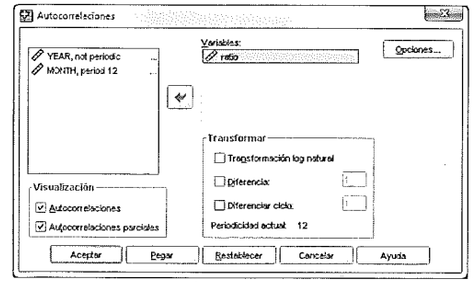


Figura 7-37

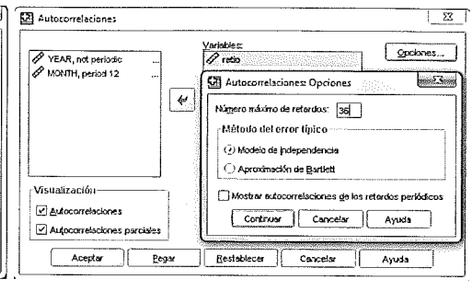


Figura 7-38

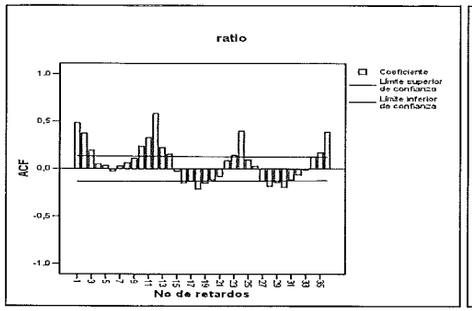


Figura 7-39

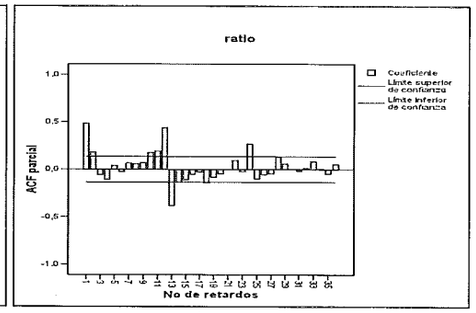


Figura 7-40

Se observa que las funciones de autocorrelación y autocorrelación parcial estimadas también validan los períodos estacionales porque los coeficientes de la FAC para retardos múltiplos del período estacional de la serie son significativamente distintos de cero. Además, para una cantidad grande de retardos la FAC se configura en forma de abanico que completa su ciclo girando sobre el eje de abscisas para una cantidad de retardos igual al período estacional. Por otro lado, la FAC parcial presenta estructura de coeficientes significativos para retardos periódicos (largos). La FAC y la FAC parcial deben considerarse a la vez, pues a veces intercambian sus papeles en el comportamiento estacional. Asímismo, los coeficientes de la FAC no decaen rápidamente, lo que indica falta de estacionariedad en media. Para certificar la falta de estacionariedad utilizamos *Analizar* → *Estadísticos descriptivos* → *Explorar* (Figura 7-41) y rellenamos la pantalla de entrada del procedimiento *Explorar* como se indica en la Figura 7-42 con la finalidad de calcular medias y varianzas por estaciones (años).

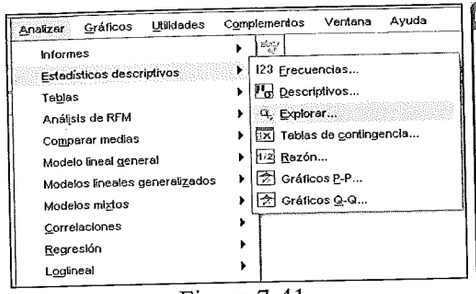


Figura 7-41

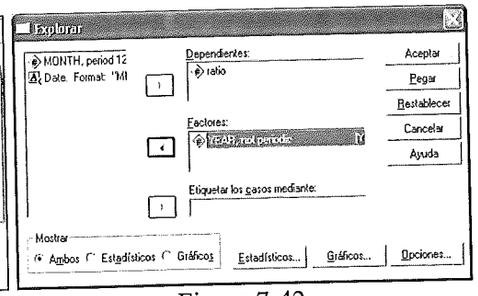


Figura 7-42

Al pulsar *Aceptar* se obtiene una sucesión de medias y varianzas por años con variaciones significativas crecientes y decrecientes a lo largo de los años, lo que nos indica que no hay estacionariedad ni en media ni en varianza en la serie original. Este resultado nos lleva a tomar logaritmos y diferenciar la serie original con el objetivo de atenuar la falta de estacionariedad en media y en varianza.

Descriptivos

	YEAR		Estadístico	Error típ.
ratio	1951	Media	1,2277	,02893
		Varianza	,010	
	1952	Media	1,2353	,03822
		Varianza	,018	
	1953	Media	1,2930	,04233
		Varianza	,022	
	1954	Media	1,3319	,03140
		Varianza	,012	
	1955	Media	1,3914	,02761
		Varianza	,009	
	1956	Media	1,4346	,02574
		Varianza	,008	
	1957	Media	1,3131	,02794
		Varianza	,009	
	1958	Media	1,1852	,02338
		Varianza	,007	
	1959	Media	1,3099	,01918
		Varianza	,004	
	1960	Media	1,2959	,03378
		Varianza	,014	
	1961	Media	1,1922	,02226
		Varianza	,006	
	1962	Media	1,2436	,02666
		Varianza	,009	
	1963	Media	1,2145	,02385
		Varianza	,007	
	1964	Media	1,2643	,03400
		Varianza	,014	
	1965	Media	1,2864	,03118
		Varianza	,012	
	1966	Media	1,2791	,03899
		Varianza	,018	
	1967	Media	1,2454	,03687
		Varianza	,016	
	1968	Media	1,2515	,04210
		Varianza	,021	
	1969	Media	1,0997	,08457
		Varianza	,014	

Pero, una vez aplicados los logaritmos, como la serie es estacional, el problema es identificar si diferenciamos la parte regular de la serie o en la parte estacional. Para ello representamos las funciones de autocorrelación estimada y autocorrelación parcial estimada bajo los supuestos de diferenciación en la parte regular (pantalla *Autocorrelaciones* según la Figura 7-43) o en la parte estacional (en la Figura 7-44 se usa 1 en *Diferenciar ciclo*).

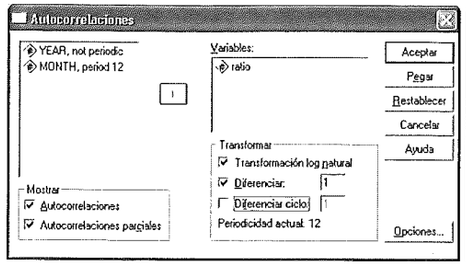


Figura 7-43

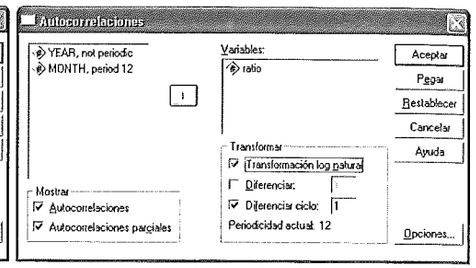


Figura 7-44

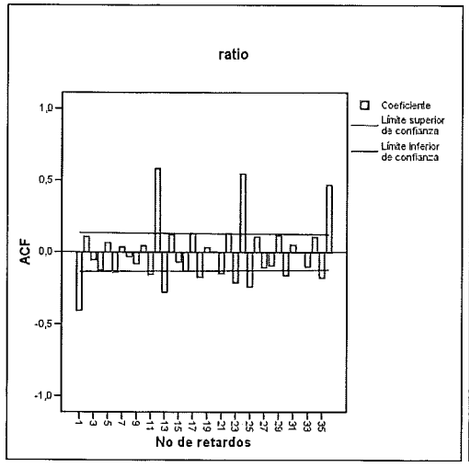


Figura 7-45

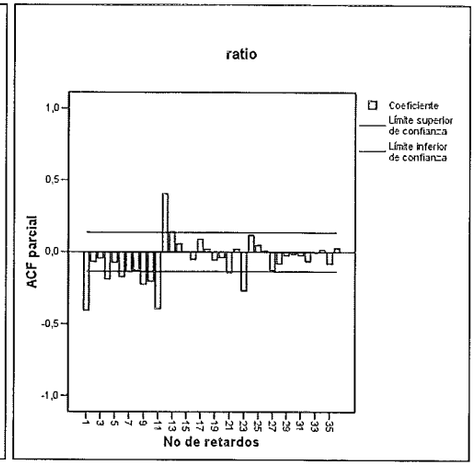


Figura 7-46

Se observa que al diferenciar sólo la parte regular, las funciones de autocorrelación (Figura 7-45) y autocorrelación parcial estimadas (Figura 7-46) no superan el problema de la falta de estacionariedad ya que la FAC no decae rápidamente. Pero al diferenciar sólo una vez la parte estacional, las funciones de autocorrelación (Figura 7-47) y autocorrelación parcial estimadas (Figura 7-48) ya superan el problema de la no estacionariedad. Asimismo, estas dos funciones cumplen las condiciones para que haya estacionalidad porque los coeficientes de la FAC para retardos múltiplos del período estacional de la serie son significativamente distintos de cero. Además, para una cantidad grande de retardos, la FAC se configura en forma de abanico que completa su ciclo girando sobre el eje de abscisas para una cantidad de retardos igual al período estacional.

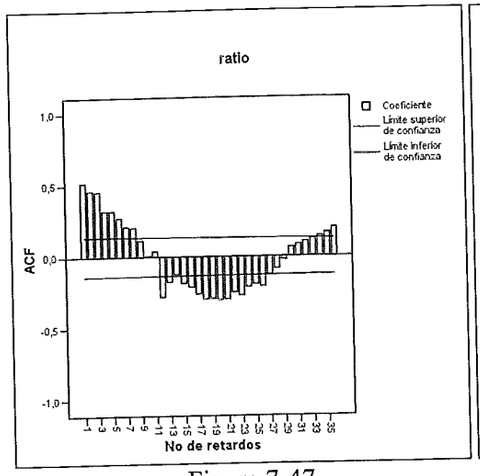


Figura 7-47

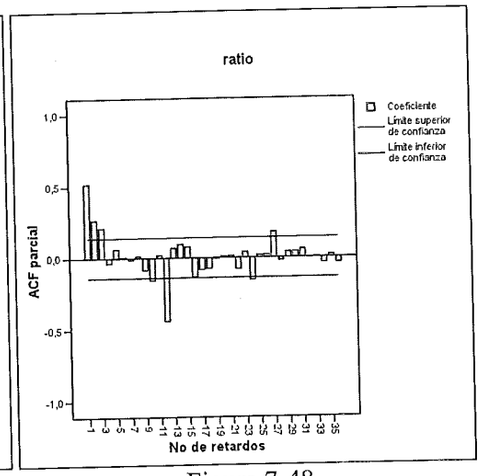


Figura 7-48

Luego el problema de la estacionalidad y la estacionariedad en media y en varianza se ha arreglado aplicando logaritmos, diferenciando una vez la parte estacional y no diferenciando la parte regular. Luego la parte regular de la serie en logaritmos es integrada de orden cero $I(0)$ y la parte estacional es integrada de orden uno $I(1)$.

Ahora resta por identificar el orden de la parte autorregresiva AR y la parte de medias móviles MA. Para realizar esta tarea utilizaremos la FAC y la FAC parcial para las que se obtuvo la estacionariedad y la estacionalidad, es decir las representadas en las Figuras 7-32 y 7-33. Observando estas dos funciones vemos que sus coeficientes no se anulan bruscamente con periodicidades y que sus estructuras se ajustan claramente a un modelo ARMA(1,1)(0,1)$_{12}$.

Existe una amplia literatura con baterías de gráficos de FAC y FAC parcial ajustadas a sus correspondientes modelos ARMA. En nuestro caso, la parte AR(1) de la parte regular proviene del decrecimiento rápido inicial y las ondas sinusoidales de la FAC añadido a que la FAC parcial presenta sólo un coeficiente significativo en la mayoría de los períodos (salvo en el primero), anulándose bruscamente el resto de los coeficientes. Asimismo, la parte MA(1) de la parte regular proviene de que la FAC presenta un sólo retardo significativo en la mayoría de los períodos (salvo en el primero). De todas formas, la única duda posible sería considerar también AR(1) la parte estacional.

Ya tenemos identificada completamente la serie inicial en logaritmos como un modelo ARIMA(1,0,1)(0,1,1)$_{12}$. Es decir, ya hemos realizado el trabajo más importante en la modelización de una serie temporal mediante la metodología de Box-Jenkins.

Una vez identificado un modelo para la serie temporal, SPSS dispone del modelizador al que se accede mediante *Analizar* → *Predicciones* → *Crear modelo* (Figura 7-49) para realizar su estimación y validación así como para obtener predicciones con la serie.

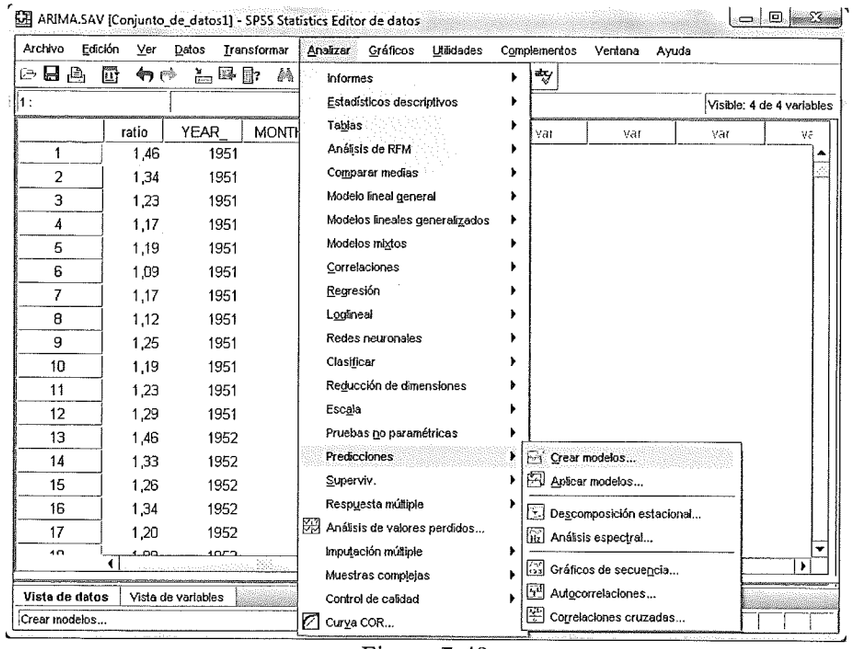


Figura 7-49

En la pantalla de entrada del modelizador seleccionamos la serie *ratio* y la desplazamos al campo *Variables dependientes*. Como ya habíamos definido fechas anteriormente parea la serie estacional *ratio*, el modelizador no pregunta por esta cuestión. El campo *Variable independiente*, que sería el tiempo, se mantiene vacío, ya que se suponen por defecto períodos de tiempo consecutivos. A continuación hacemos clic en la flecha situada a la derecha de la opción *Modelizador experto* del campo *Método* y elegimos *ARIMA* (Figura 7-50). El modelizador sitúa ya la opción de trabajo con modelos ARIMA (Figura 7-51). A continuación hacemos clic en el botón *Criterios* y en la pantalla *Criterios de ARIMA* (Figura 7-52) introducimos los parámetros del modelo ARIMA identificado previamente como ARIMA(1,0,1)(0,1,1)$_{12}$ para la serie en logaritmos. Al hacer clic en *Continuar*, ya tenemos la pestaña *Variables* del modelizador rellenada adecuadamente.

A continuación hacemos clic en la pestaña *Estadísticos* y elegimos los estadísticos deseados en la salida (Figura 7-53. En la pestaña *Gráficos* se eligen las opciones gráficas deseadas en la salida (Figura 7-54).

A continuación hacemos clic en la pestaña *Filtro de resultados* para incluir todos los modelos posibles en los resultados (Figura 7-55). La siguiente tarea es hacer clic en la pestaña *Guardar* para elegir las series a guardar (Figura 7-56). Por último hacemos clic en la pestaña *Opciones* y situamos la fecha final para las predicciones en el campo *Fecha* (Figura 7-57) para predecir datos hasta el último mes de 1969. Previamente se ha marcado la opción para situar el *Periodo de prediccion* desde el *Primer caso después del final del período de estimación hasta una fecha especificada.*

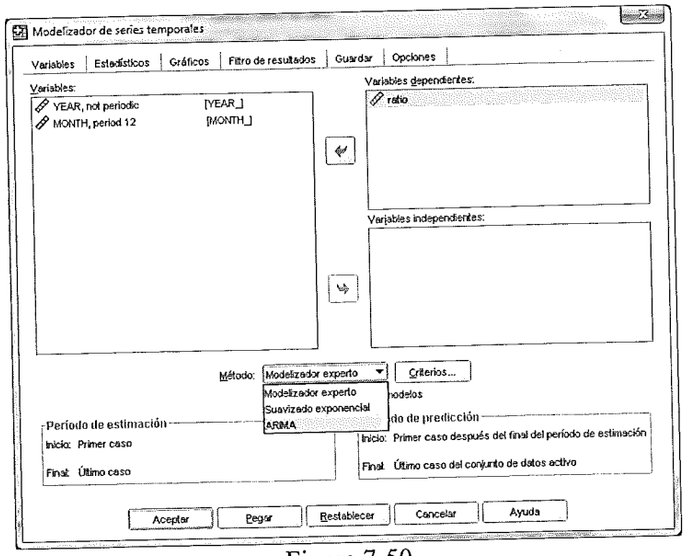


Figura 7-50

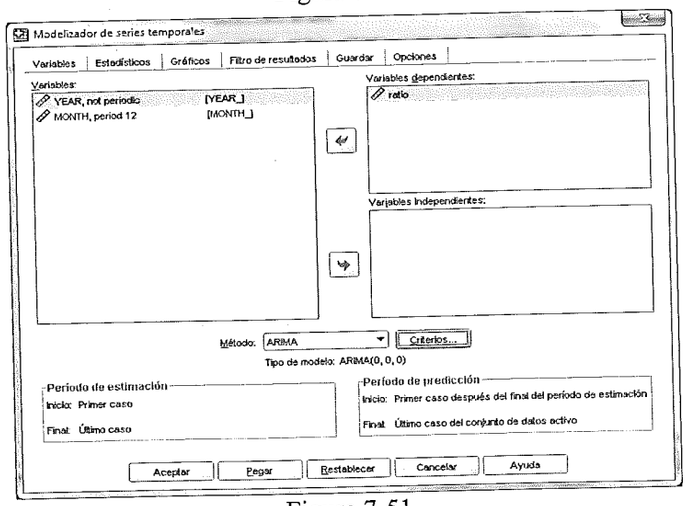


Figura 7-51

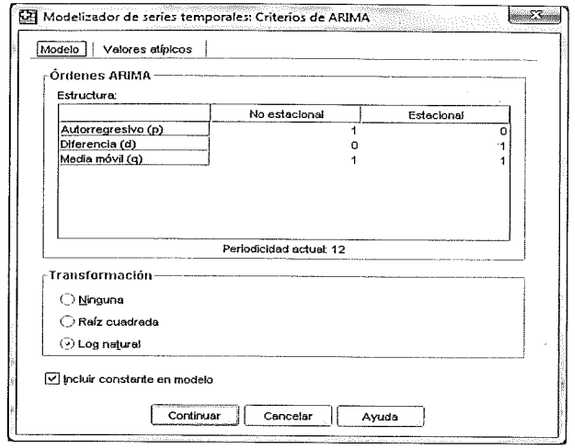


Figura 7-52

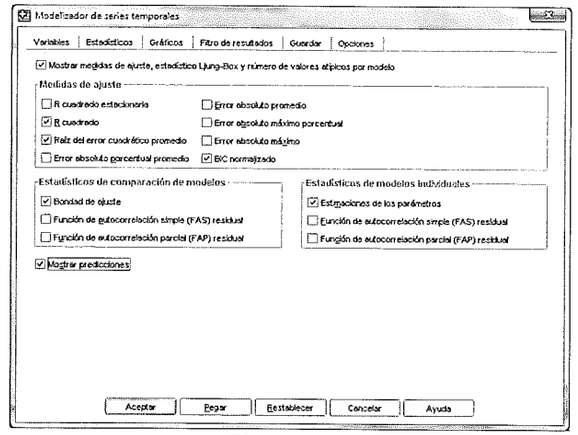


Figura 7-53

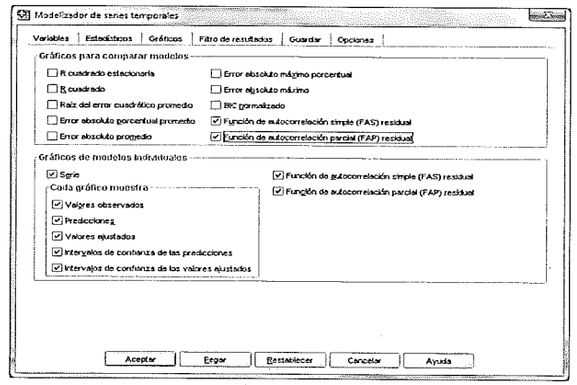


Figura 7-54

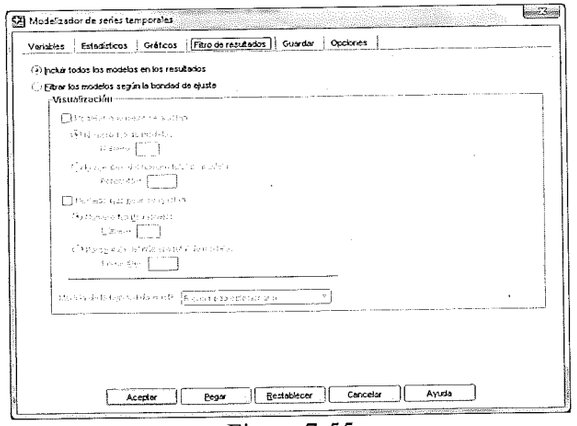


Figura 7-55

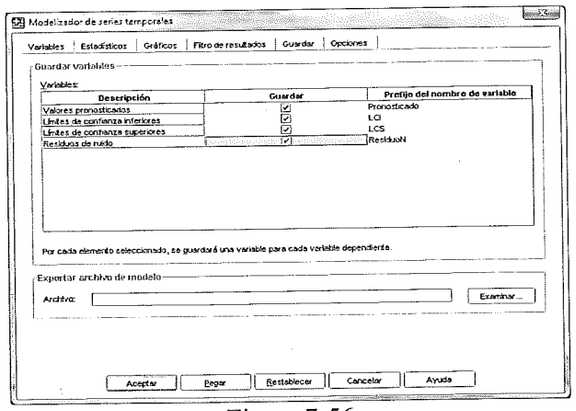


Figura 7-56

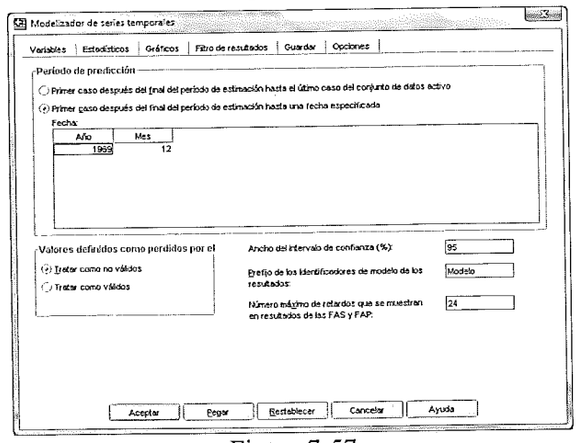


Figura 7-57

Al hacer clic en *Aceptar* se obtienen los estadísticos de ajuste del modelo (Figura 7-58). No obstante, se observa que el p-valor correspondiente a la constante en la tabla ARIMA *Model Parameters* es muy alto (0,761), lo que invalida su presencia en el modelo por tener muy baja significatividad. Este hecho conduce a eliminar la citada constante del modelo de ajuste de la serie. Para ello, en la pantalla *Criterios de ARIMA* de la Figura 7-52 se elimina la marca de la casilla *Incluir constante en modelo* (Figura 7-59) y se ejecuta nuevamente la estimación con el modelizador. Ahora, los resultados de la estimación de la Figura 7-60 son correctos, ya que los parámetros estimados tienen una significatividad muy alta (p-valores muy pequeños). En la salida también se observan otros estadísticos de ajuste como un R^2 de valor 0,651 y un p-valor del estadístico de Ljung-Box mayor que 0,05 que indica aleatoriedad de los residuos al 95%. La gráfica de la función de autocorrelación residual (Figura 7-61) corrobora este hecho. En la salida también se observan las predicciones, ya que había sido marcado el campo *Mostrar predicciones* en la pantalla *Estadísticos* de la Figura 7-53. Asimismo, el fichero de datos presenta una columna adicional a la derecha con las predicciones (Figura 7-62). Por último, se observa la gráfica de la serie suavizada, las predicciones y sus intervalos de confianza (Figura 7-63).

ARIMA Model Parameters

				Estimate	SE	t	Sig.
ratio-Modelo_1	ratio	Natural Log	Constant	-,002	,008	-,305	,761
			AR Lag 1	,903	,044	20,673	,000
			MA Lag 1	,529	,085	6,243	,000
			Seasonal Difference	1			
			MA, Seasonal Lag 1	,631	,070	9,041	,000

Forecast

Model		Mar 1969	Apr 1969	May 1969	Jun 1969	Jul 1969	Aug 1969	Sep 1969	Oct 1969	Nov 1969	Dec 1969
ratio-Modelo_1	Forecast	1,22	1,11	1,13	,99	1,13	1,12	1,46	1,26	1,21	1,30
	UCL	1,37	1,24	1,28	1,13	1,28	1,28	1,68	1,44	1,39	1,50
	LCL	1,09	,98	,99	,87	,98	,97	1,27	1,09	1,04	1,13

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

Figura 7-58

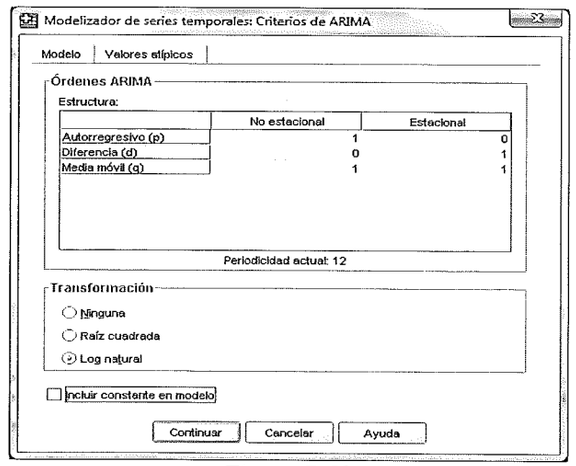


Figura 7-59

Model Statistics

Model	Number of Predictors	Model Fit statistics			Ljung-Box Q(18)			Number of Outliers
		R-squared	RMSE	Normalized BIC	Statistics	DF	Sig.	
ratio-Modelo_1	0	,651	,074	-5,141	22,303	15	,100	0

ARIMA Model Parameters

					Estimate	SE	t	Sig.
ratio-Modelo_1	ratio	Natural Log	AR	Lag 1	,905	,043	20,924	,000
			MA	Lag 1	,531	,084	6,302	,000
			Seasonal Difference		1			
			MA, Seasonal	Lag 1	,632	,069	9,111	,000

Forecast

Model		Mar 1969	Apr 1969	May 1969	Jun 1969	Jul 1969	Aug 1969	Sep 1969	Oct 1969	Nov 1969	Dec 1969
ratio-Modelo_1	Forecast	1,22	1,11	1,13	,99	1,13	1,12	1,47	1,26	1,21	1,31
	UCL	1,37	1,25	1,28	1,13	1,29	1,29	1,68	1,45	1,40	1,51
	LCL	1,09	,98	,99	,87	,98	,98	1,27	1,09	1,05	1,13

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

Figura 7-60

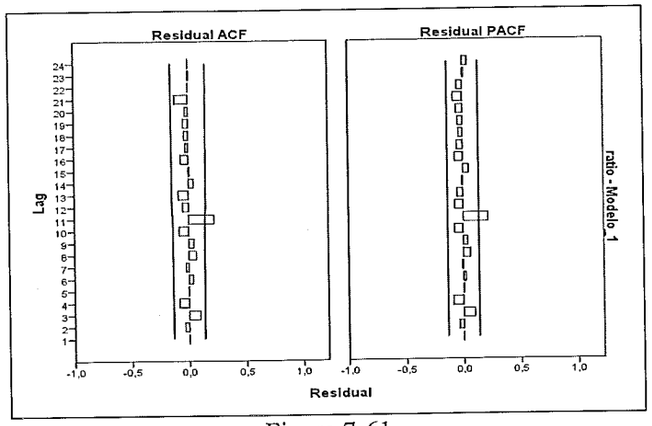


Figura 7-61

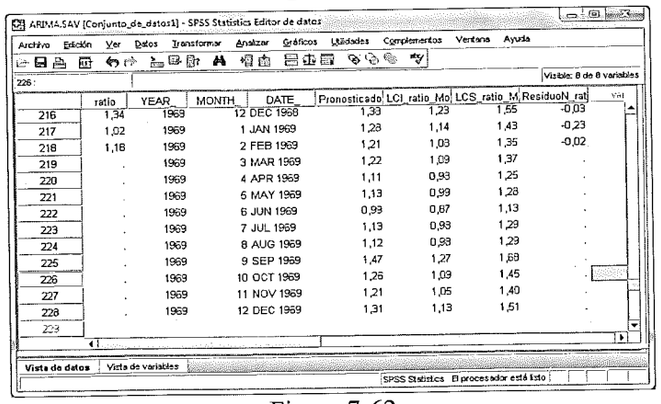


Figura 7-62

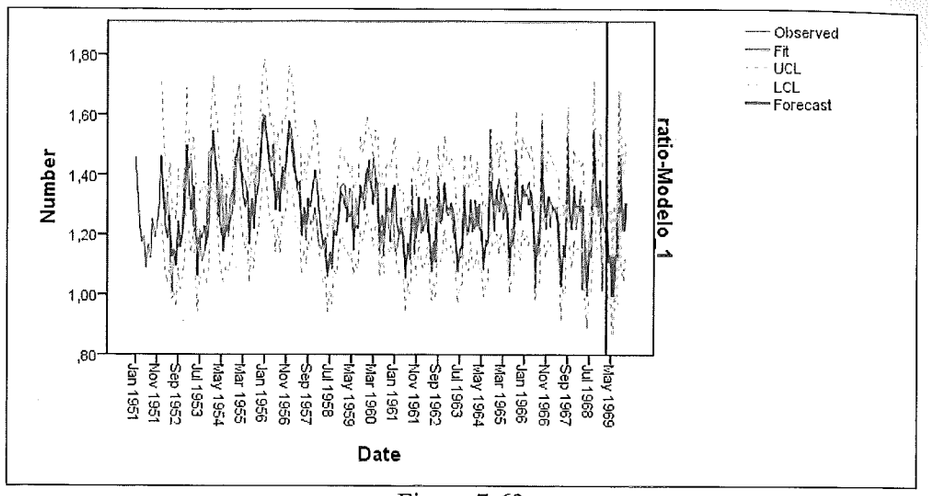


Figura 7-63

PREDICCIONES INCONDICIONALES DETERMINISTAS CON EL MODELIZADOR DE SPSS. SUAVIZADO

SPSS realiza el suavizado de series temporales mediante métodos deterministas a través del procedimiento *Suavizado exponencial*. Este procedimiento suaviza componentes irregulares de datos de series temporales, para lo cual hace uso de una variedad de modelos que incorporan diferentes supuestos acerca de la tendencia y la estacionalidad. Como ejemplo consideramos la serie temporal de nombre *Cantidad* contenida en el archivo *suavizado.sav* y vamos a realizar un suavizado de la misma. En primer lugar la representamos (*Analizar* → *Predicciones* → *Gráfico de secuencia*) con el objeto de observar si presenta estacionalidad. Para ello se rellena la pantalla de entrada de *Gráficos de secuencia* según la Figura 7-64 y se obtiene el gráfico de la Figura 7-65.

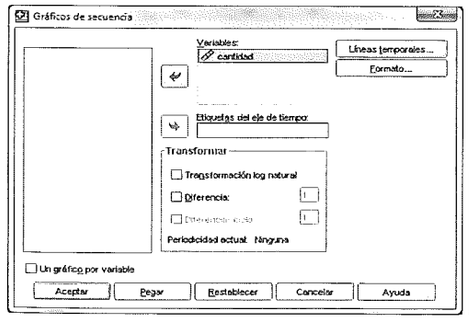


Figura 7-64

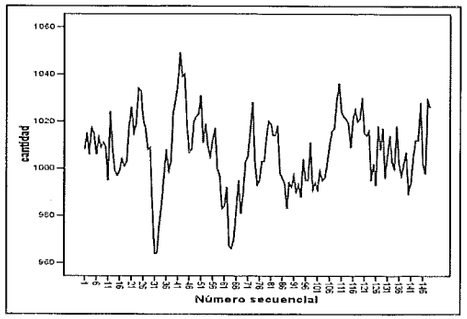


Figura 7-65

Gráficamente no se observa estacionalidad. Para obtener el periodograma de la serie usamos *Gráficos* → *Serie temporal* → *Análisis espectral* y rellenamos la pantalla de entrada del procedimiento *Diagramas espectrales* según se indica en la Figura 7-66. Al pulsar *Aceptar* se obtiene el periodograma de frecuencia de la Figura 7-67, que no presenta picos destacados, lo que hace intuir que no hay estacionalidad.

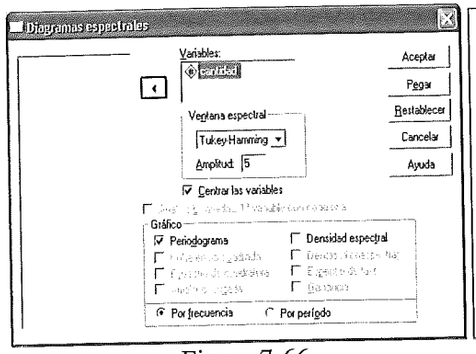


Figura 7-66

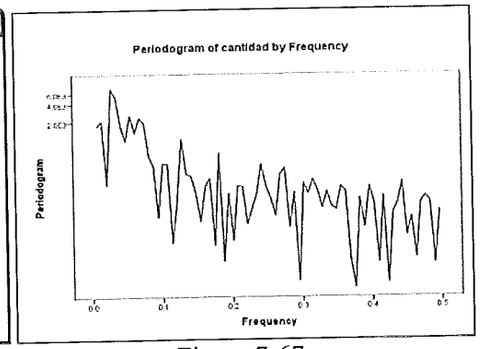


Figura 7-67

Para realizar el suavizado se utiliza el modelizador de SPSS. Para ello elija en los menús *Analizar* → *Predicciones* → *Crear modelos* (Figura 7-68). Si la serie hubiese sido estacional sería necesario definir fechas previamente. Aún así, el modelizador pregunta si es necesario definir fechas (Figura 7-60). Podemos hacer clic en *Definir fecha* y realizar la definición adecuada si fuese necesario. En nuestro caso pulsamos *Aceptar* para considerar la serie *Cantidad* como temporal tratando los casos como períodos de tiempo consecutivos. Se obtiene el *Modelizador de series temporales* de la Figura 7-70.

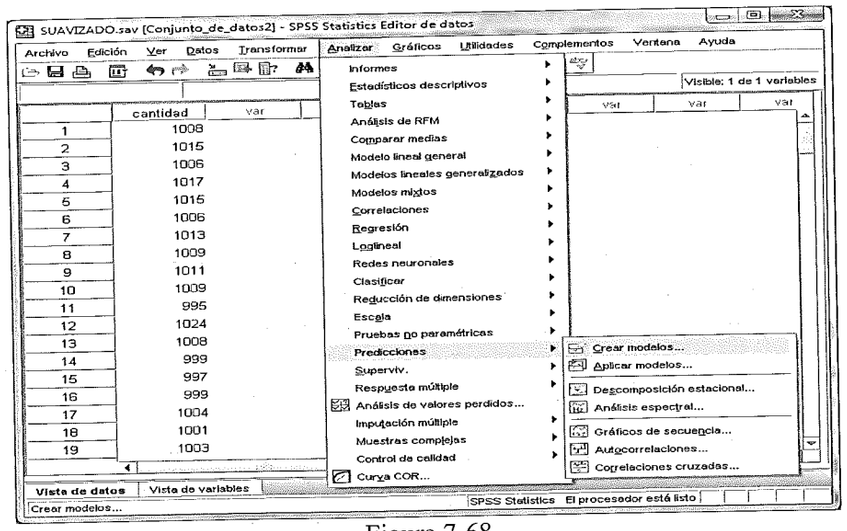


Figura 7-68

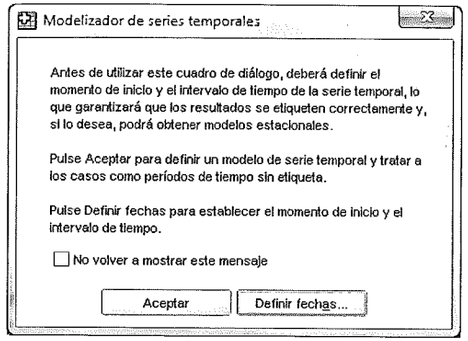


Figura 7-69

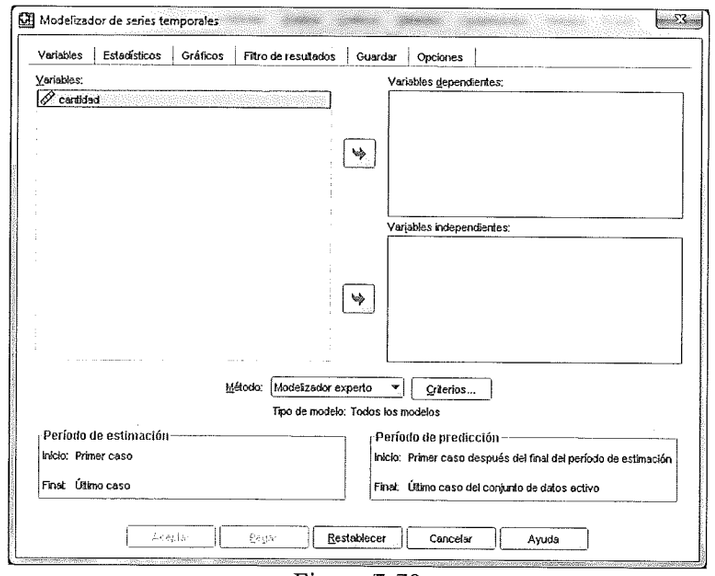


Figura 7-70

Seleccionamos la serie Cantidad y la desplazamos al campo *Variables dependientes*. El campo *Variable independiente*, que sería el tiempo, se mantiene vacío, ya que se suponen por defecto períodos de tiempo consecutivos. A continuación hacemos clic en la flecha situada ala derecha de la opción *Modelizador experto* del campo *Método* y elegimos *Suavizado exponencial* (Figura 7-71). En la pantalla *Criterios de suavizado exponencial* (Figura 7-72) elegimos uno cualquiera de ellos, por ejemplo *Tendencia lineal de Brown*. Al hacer clic en *Continuar*, ya tenemos la pestaña *Variables* del modelizador rellenada adecuadamente (Figura 7-73). A continuación hacemos clic en la pestaña *Estadísticos* y elegimos los estadísticos deseados en la salida (Figura 7-74). En la pestaña *Gráficos* se eligen las opciones gráficas deseadas en la salida (Figura 7-75).

A continuación hacemos clic en la pestaña *Filtro de resultados* para incluir todos los modelos posibles en los resultados (Figura 7-76). La siguiente tarea es hacer clic en la pestaña *Guardar* para elegir las series a guardar (Figura 7-77). Por último hacemos clic en la pestaña *Opciones* y situamos el valor 155 en la casilla *Observación* del campo *Fecha* (Figura 7-78), ya que queremos predecir 6 observaciones a partir de la número 149 (que es la última de la serie). Previamente se ha marcado la opción para situar el *Periodo de predicción* desde el *Primer caso después del final del período de estimación hasta una fecha especificada.*

Al hacer clic en *Aceptar* se obtienen los estadísticos de ajuste del modelo (Figura 7-79) entre los que destacan R^2=0,5 y el p-valor del estadístico de Lunj Box lo suficientemente alto como para asegurar la aleatoriedad residual. Finalmente se obtienen las funciones de autocorrelación y autocorrelación parcial residuales (Figura 7-80) cuyos valores no se salen de las bandas de confianza, lo que certifica una buena diagnosis residual. Se obtiene también sobre la misma gráfica la serie original, la serie suavizada y las predicciones obtenidas (Figura 7-81). Si observamos el conjunto de datos vemos que al final se han añadido las series de predicciones, sus intervalos de confianza y la serie de errores (Figura 7-82).

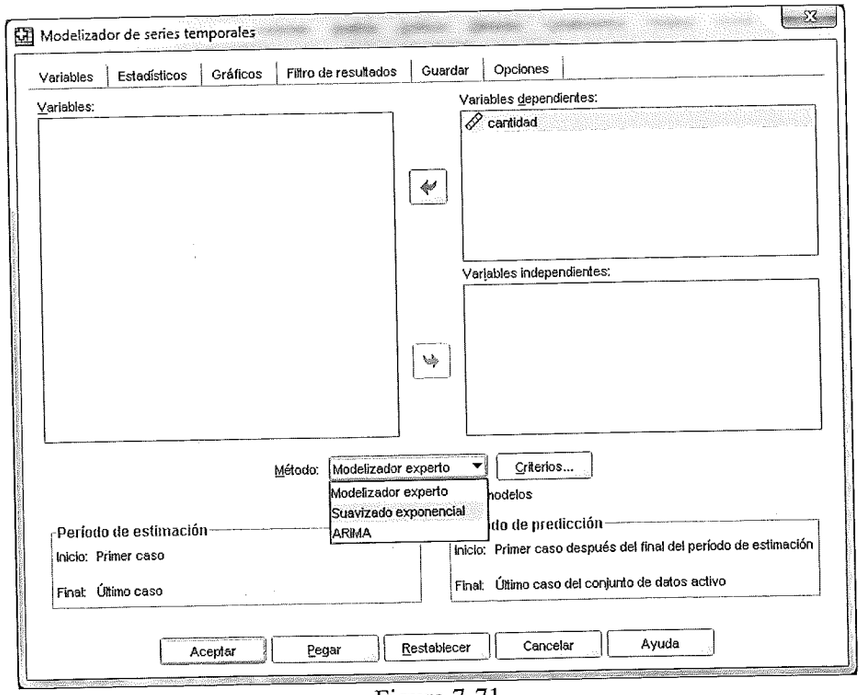


Figura 7-71

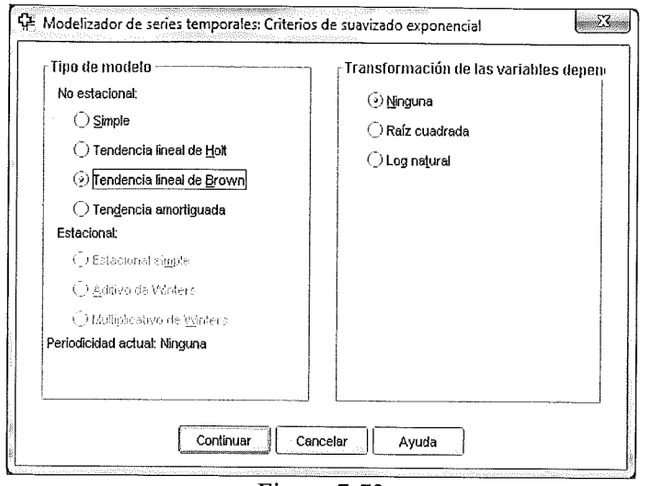


Figura 7-72

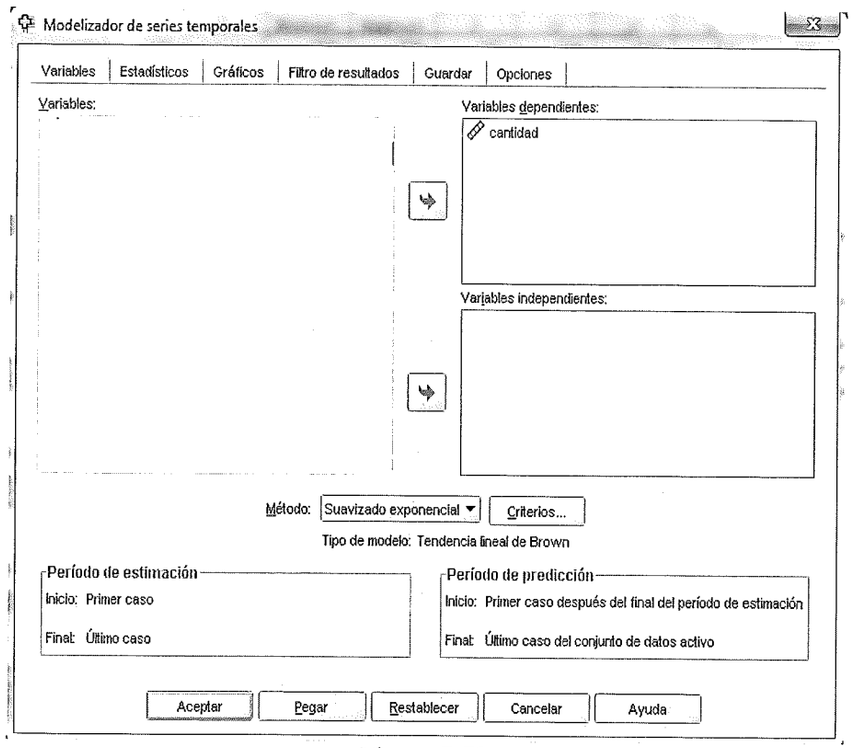


Figura 7-73

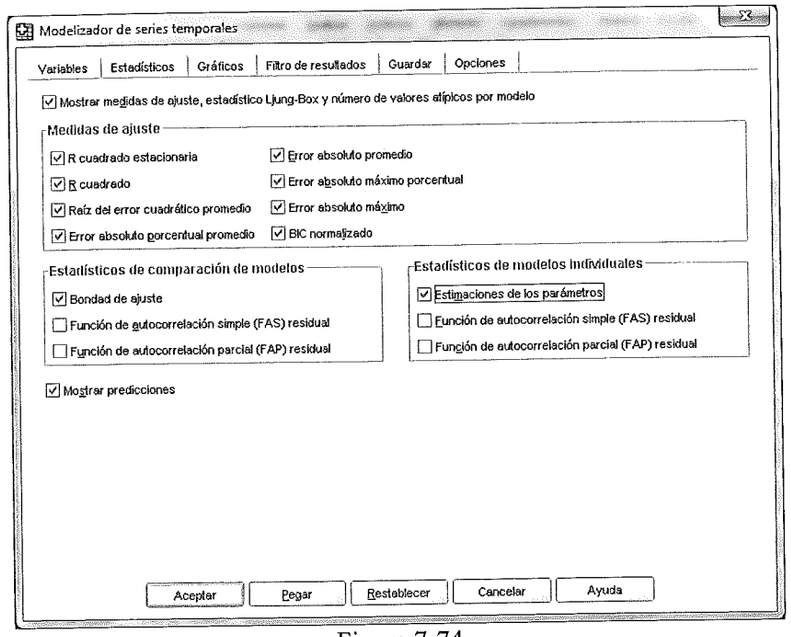


Figura 7-74

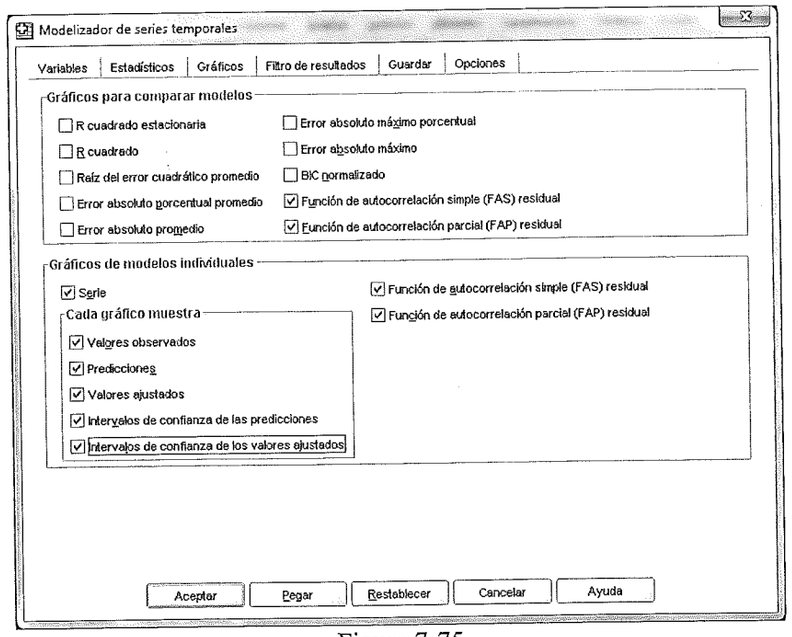


Figura 7-75

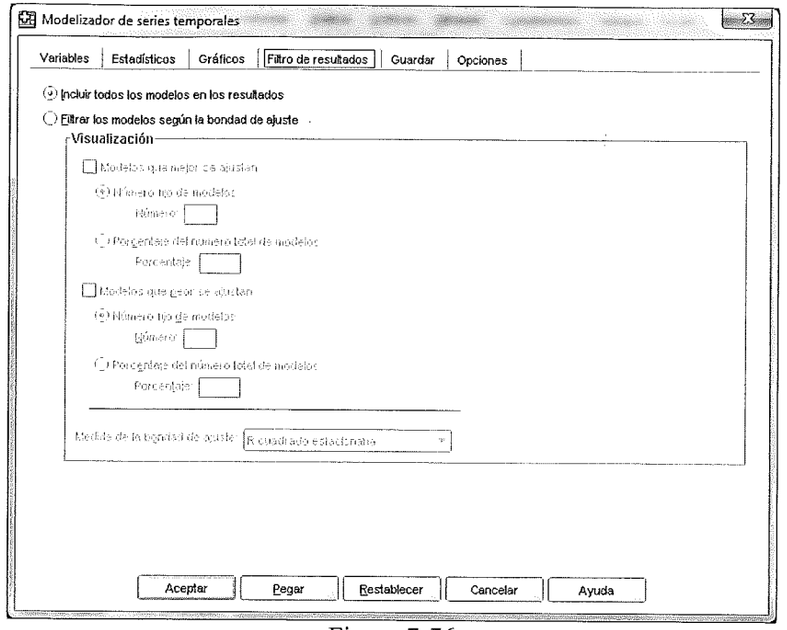


Figura 7-76

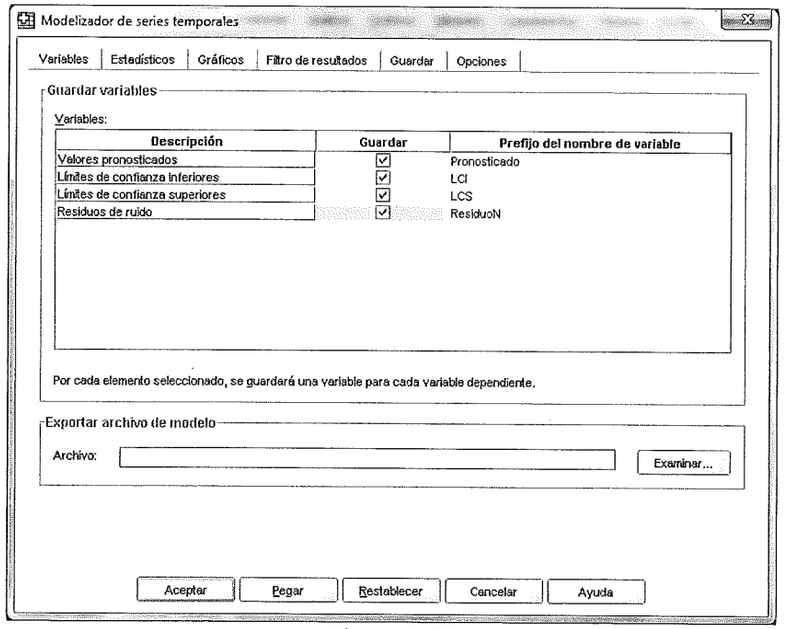


Figura 7-77

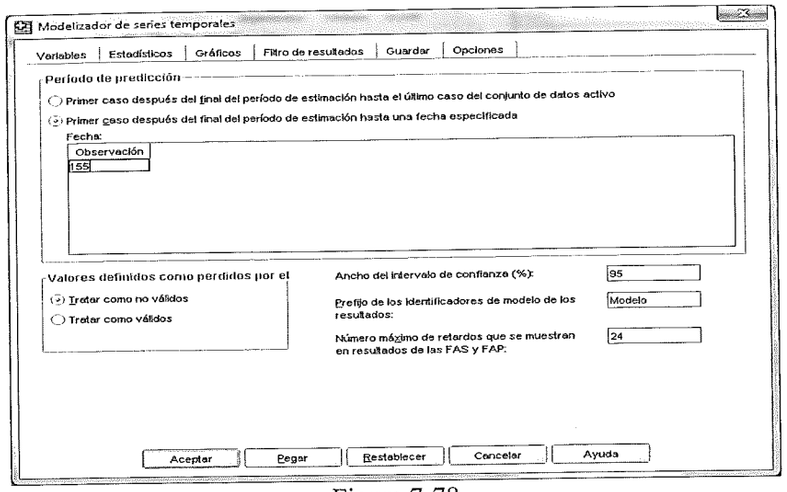


Figura 7-78

Model Statistics

Model	Number of Predictors	Model Fit statistics: Stationary R-squared	R-squared	RMSE	MAPE	MAE	MaxAPE	MaxAE	Normalized BIC	Ljung-Box Q(18): Statistics	DF	Sig.	Number of Outliers
cantidad-Modelo_1	0	,509	,438	11,760	,931	9,363	2,949	29,583	4,963	19,373	17	,308	0

Exponential Smoothing Model Parameters

Model			Estimate	SE	t	Sig.
cantidad-Modelo_1	No Transformation	Alpha (Level and Trend)	,427	,034	12,612	,000

Forecast

Model		150	151	152	153	154	155
cantidad-Modelo_1	Forecast	1029	1032	1036	1040	1044	1047
	UCL	1052	1063	1075	1088	1102	1116
	LCL	1005	1002	997	992	986	979

Figura 7-79

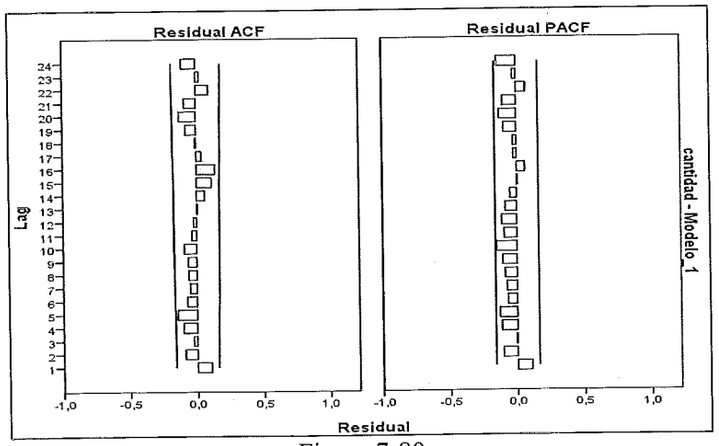


Figura 7-80

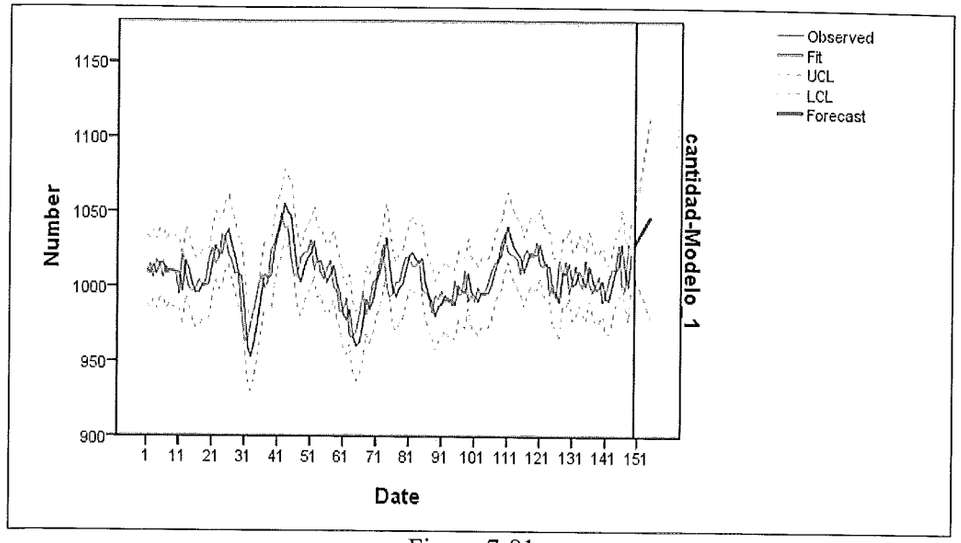
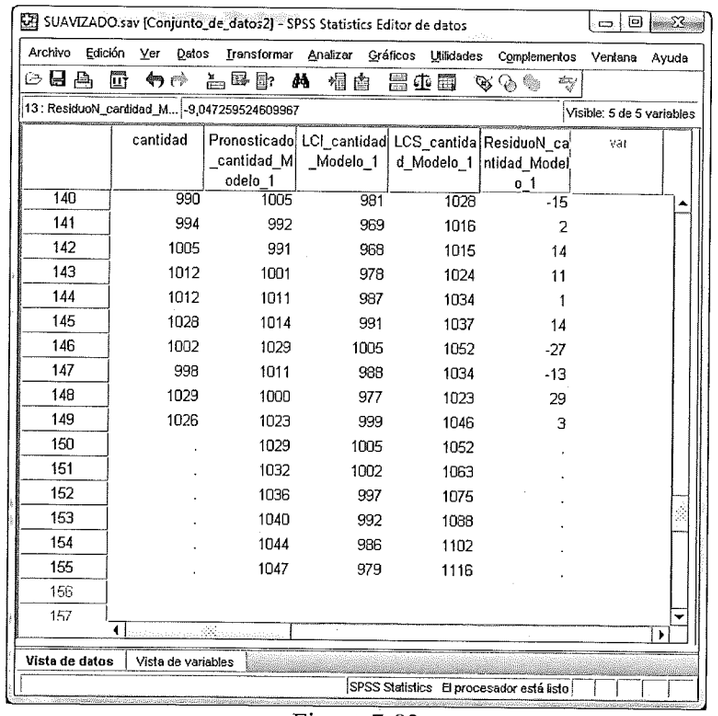


Figura 7-81



	cantidad	Pronosticado_cantidad_Modelo_1	LCI_cantidad_Modelo_1	LCS_cantidad_Modelo_1	ResiduoN_cantidad_Modelo_1	var
140	990	1005	981	1028	-15	
141	994	992	969	1016	2	
142	1005	991	968	1015	14	
143	1012	1001	978	1024	11	
144	1012	1011	987	1034	1	
145	1028	1014	991	1037	14	
146	1002	1029	1005	1052	-27	
147	998	1011	988	1034	-13	
148	1029	1000	977	1023	29	
149	1026	1023	999	1046	3	
150	.	1029	1005	1052	.	
151	.	1032	1002	1063	.	
152	.	1036	997	1075	.	
153	.	1040	992	1088	.	
154	.	1044	986	1102	.	
155	.	1047	979	1116	.	
156						
157						

Figura 7-82

SPSS Y LOS MODELOS ARIMA CON INTERVENCIÓN

El modelizador automático de SPSS incorpora también el análisis de la intervención. Como ejemplo consideramos la serie mensual del impuesto sobre sociedades *is00.sav* ingresado por la Administración tributaria a lo largo de varios años. Se trata de predecir ingresos futuros por este impuesto incorporando análisis de la intervención.

Comenzamos abriendo el fichero *is00.sav* a través de la opción *Abrir Datos* del menú *Archivo*. Para comenzar la fase de identificación de la serie utilizamos *Analizar* → *Predicciones* → *Crear modelos* con el objeto de iniciar el modelizador.

Como no hay variable fecha definida en el fichero en formato SPSS, obtenemos la Figura 7-83, cuyo botón *Definir fechas* nos lleva a la Figura 7-84 para indicar el primer caso de la serie. Al hacer clic en *Aceptar* se obtiene la Figura 7-85 y las series ya están fechadas. Situamos la variable *is00* en el campo *Variables dependientes* y en el campo *Método* de la pantalla inicial del *Modelizador de series temporales* se elige *Modelizador experto* (Figura 7-85).

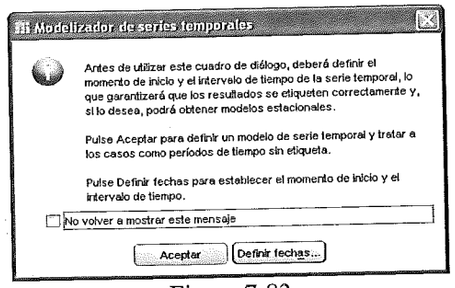


Figura 7-83

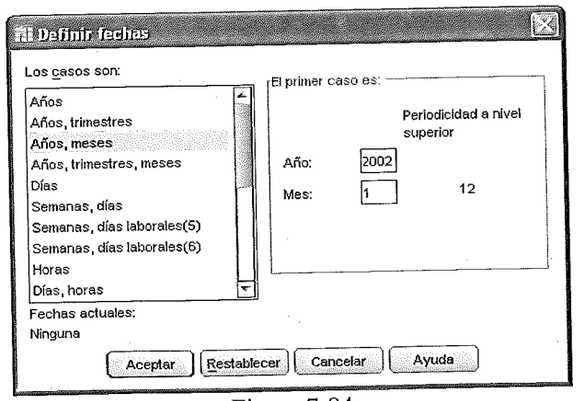


Figura 7-84

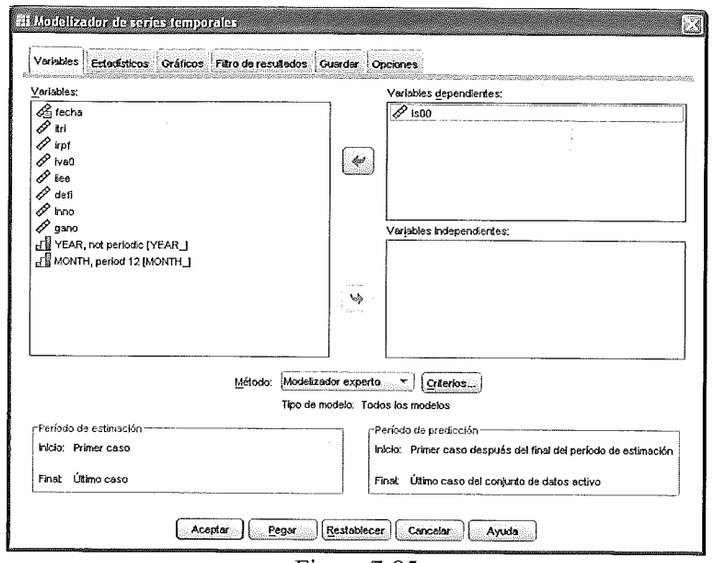


Figura 7-85

Se hace clic en *Criterios* y en la pantalla *Criterios del modelizador experto* se elige *Sólo modelos ARIMA* en la solapa *Modelo* (Figura 7-86). En la solapa *Valores atípicos* se marca toda la tipología de intervenciones (Figura 7-87). Se hace clic en *Continuar* y ya se tiene cumplimentada la pestaña *Variables* del modelizador (Figura 7-88). El resto de las pestañas del modelizador se rellenan como se indica en las Figuras 7-89 a 7-93.

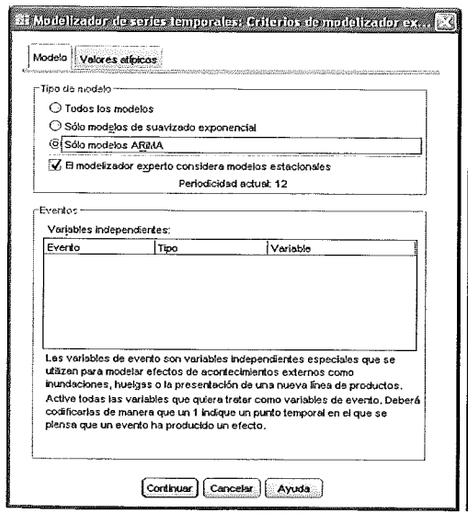


Figura 7-86

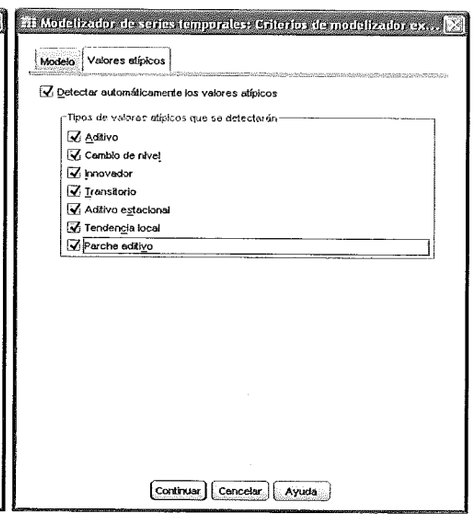


Figura 7-87

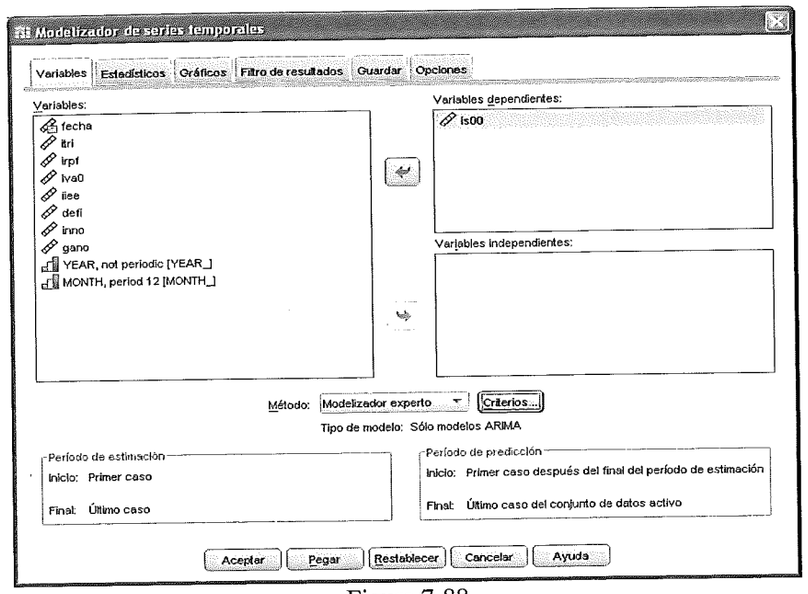


Figura 7-88

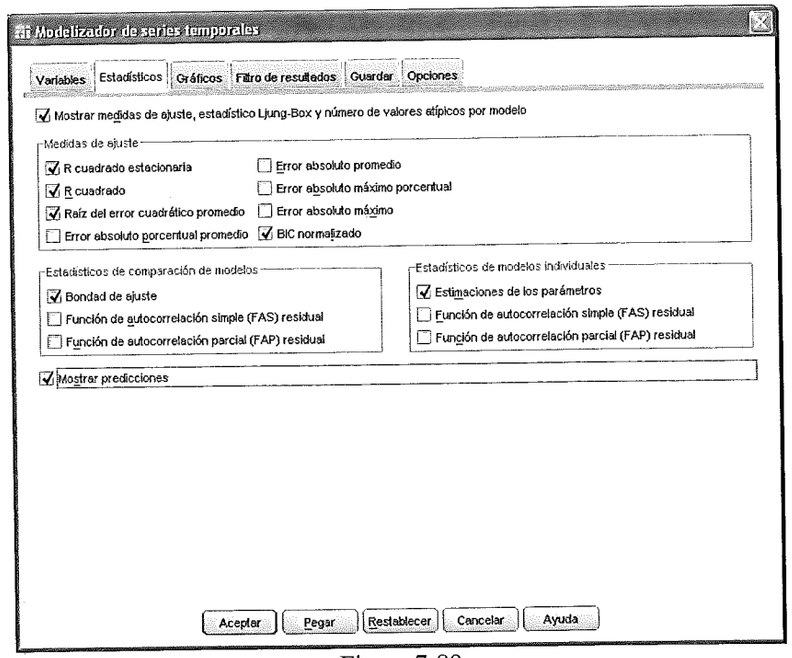


Figura 7-89

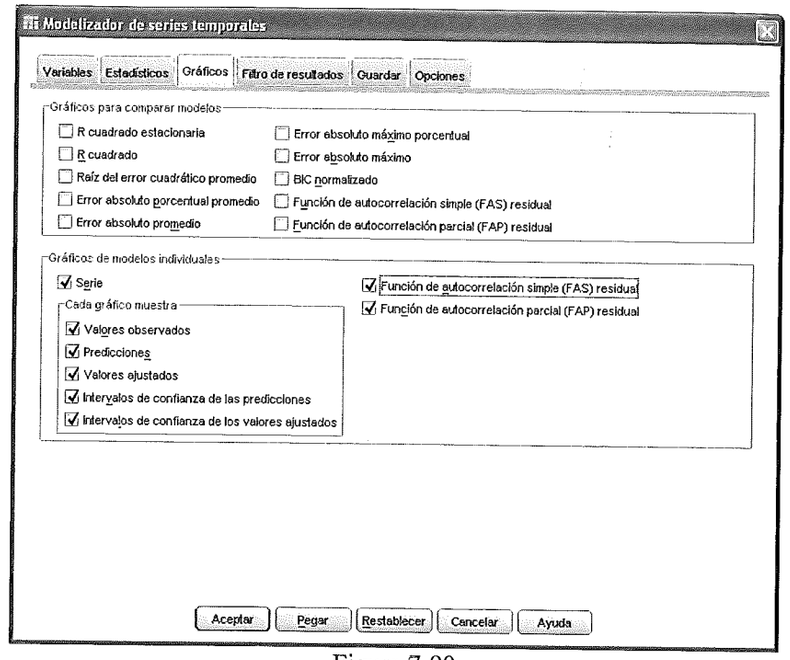


Figura 7-90

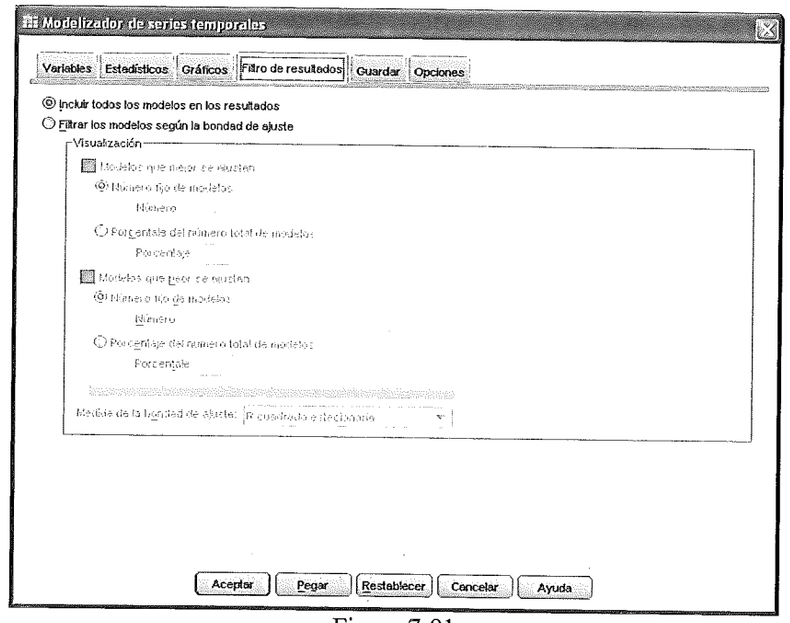


Figura 7-91

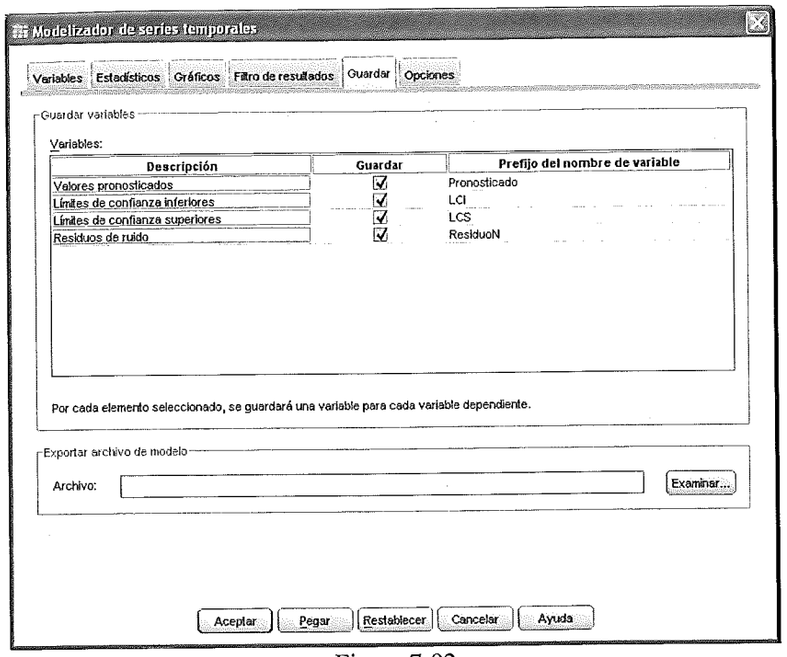


Figura 7-92

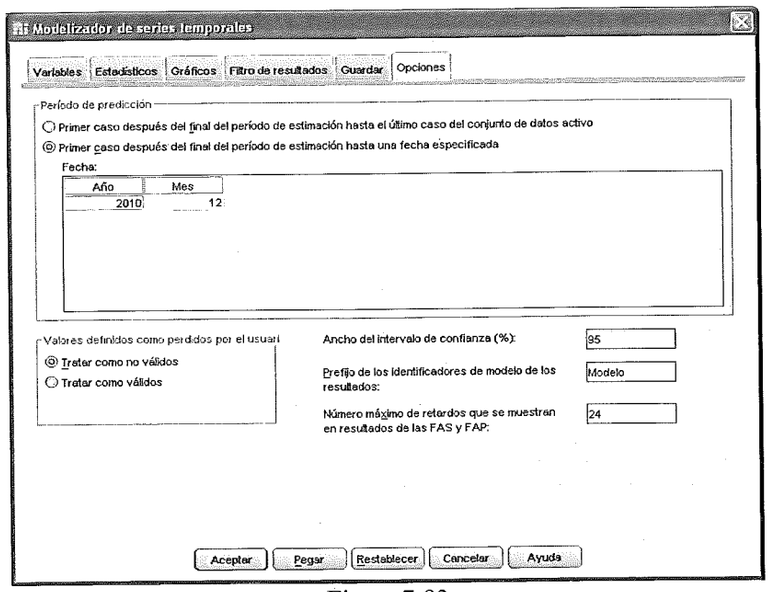


Figura 7-93

Al hacer clic en *Aceptar*, se observa en la Figura 7-94 que muestra el modelo ARIMA elegido automáticamente por el modelizador. Se trata de un modelo estacional puro ARIMA(0,0,0)(0,1,1)$_{12}$. Los resultados del modelo estimado y su diagnosis se observan en la Figura 7-95. Se observa la alta significatividad de los parámetros estimados (p-valores casi nulos) y el estadístico de Lunj Box con p-valor alto (0,335) que certifica la aleatoriedad residual. Se observan cuatro intervenciones (dos outliers aditivos AO, dos outliers innovadores IO y un cambio de nivel) en las fechas indicadas en la Figura 7-95. Se observa que las intervenciones son muy significativas (p-valores muy pequeños). En la Figura 7-96 se observan las predicciones pedidas (hasta final del año 2010).

Modelizador para series temporales

```
[Conjunto_de_datos1] C:\libros\SERIESISABEL\taxes.sav
```

Descripción del modelo

			Tipo de modelo
ID del modelo	is00	Modelo_1	ARIMA(0,0,0)(0,1,1)

Resumen del modelo

Ajuste del modelo

Estadístico de ajuste	Media	ET	Mínimo	Máximo	Percentil 5	10	25	50	75	90	95
R-cuadrado estacionaria	,803	.	,803	,803	,803	,803	,803	,803	,803	,803	,803
R-cuadrado	,921	.	,921	,921	,921	,921	,921	,921	,921	,921	,921
RMSE	1061,925	.	1061,925	1061,925	1061,925	1061,925	1061,925	1061,925	1061,925	1061,925	1061,925
MAPE	17,055	.	17,055	17,055	17,055	17,055	17,055	17,055	17,055	17,055	17,055
MaxAPE	78,267	.	78,267	78,267	78,267	78,267	78,267	78,267	78,267	78,267	78,267
MAE	437,618	.	437,618	437,618	437,618	437,618	437,618	437,618	437,618	437,618	437,618
MaxAE	6286,127	.	6286,127	6286,127	6286,127	6286,127	6286,127	6286,127	6286,127	6286,127	6286,127
BIC normalizado	14,289	.	14,289	14,289	14,289	14,289	14,289	14,289	14,289	14,289	14,289

Figura 7-94

Estadísticos del modelo

Modelo	Número de predictores	Estadísticos de ajuste del modelo: R-cuadrado estacionaria	R-cuadrado	RMSE	BIC normalizado	Ljung-Box Q(18): Estadísticos	GL	Sig.	Número de valores atípicos
is00-Modelo_1	0	,803	,921	1061,925	14,289	18,895	17	,335	5

Parámetros del modelo ARIMA

				Estimación	ET	t	Sig.
is00-Modelo_1	is00	Log natural	Constante	,106	,015	7,014	,000
			Diferenciación estacional	1			
			MA, estacional Retardo 1	,465	,093	4,996	,000

Valores atípicos

			Estimación	ET	t	Sig.
is00-Modelo_1	Ene 2004	Aditivo	-,799	,181	-4,405	,000
	Abr 2008	Innovador	-1,395	,213	-6,556	,000
	May 2008	Aditivo	2,150	,182	11,816	,000
	Oct 2008	Innovador	-,752	,212	-3,543	,001
	Jul 2009	Cambio de nivel	-,392	,071	-5,486	,000

Figura 7-95

Previsión

Modelo		Jun 2010	Jul 2010	Ago 2010	Sep 2010	Oct 2010	Nov 2010	Dic 2010
is00-Modelo_1	Previsión	667,2	1823,7	10085,6	690,6	7634,1	609,6	3859,9
	LCS	991,7	2710,7	14990,9	1026,4	11347,1	906,0	5737,2
	LCI	429,5	1173,8	6491,7	444,5	4913,8	392,4	2484,5

Para cada modelo, las predicciones comienzan después del último valor no perdido del rango del período de estimación solicitado y finalizan en el último período para el que hay disponibles valores no perdidos de todos los predictores o en la fecha de finalización del período de predicción solicitado, lo que ocurra antes.

Figura 7-96

Los gráficos de las funciones de autocorrelación residuales de la Figura 7-97 indican que los residuos son aleatorios (valores dentro de las franjas de confianza). La Figura 7-98 muestra la serie suavizada con predicciones e intervalos de confianza. La Figura 7-99 muestra las predicciones y sus intervalos de confianza guardadas como variables a la derecha del fichero de datos.

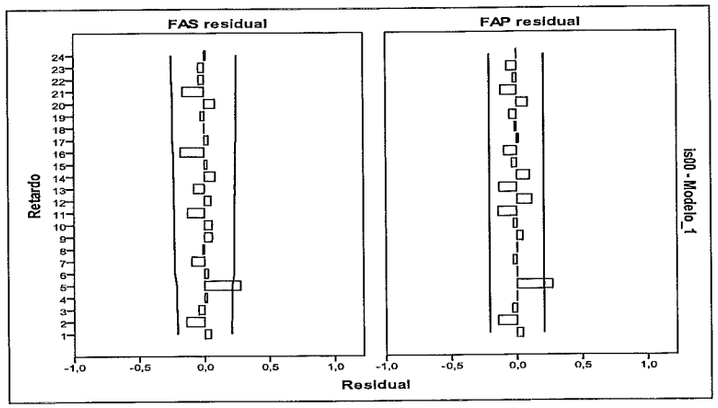


Figura 7-97

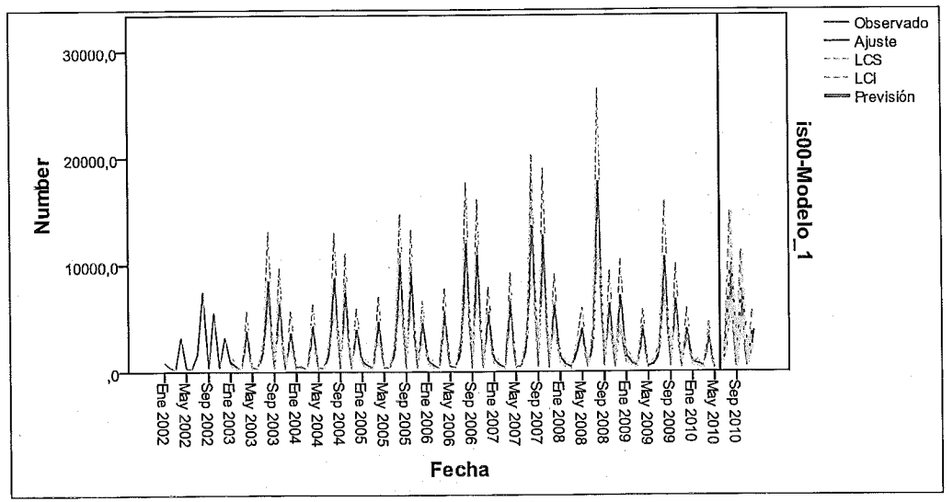


Figura 7-98

	YEAR_	MONTH_	DATE_	Pronosticado_is00_Modelo_1	LCI_is00_Modelo_1	LCS_is00_Modelo_1	ResiduoN_is00_Modelo_1
99	2010	3	MAR 2010	425,1	273,6	631,8	-,2
100	2010	4	APR 2010	3121,7	2009,3	4639,9	,2
101	2010	5	MAY 2010	349,1	224,7	518,9	,0
102	2010	6	JUN 2010	667,2	429,5	991,7	.
103	2010	7	JUL 2010	1823,7	1173,8	2710,7	.
104	2010	8	AUG 2010	10085,6	6491,7	14990,9	.
105	2010	9	SEP 2010	690,6	444,5	1026,4	.
106	2010	10	OCT 2010	7634,1	4913,8	11347,1	.
107	2010	11	NOV 2010	609,6	392,4	906,0	.
108	2010	12	DEC 2010	3859,9	2484,5	5737,2	.

Figura 7-99

SAS Y LA IDENTIFICACIÓN, ESTIMACIÓN, VALIDACIÓN Y PREDICCIÓN DE MODELOS ARIMA(*p,d,q*)(*P,D,Q*)*s*

Procedimiento ARIMA

La sintaxis del procedimiento ARIMA, que ajusta modelos de series temporales según la metodología de Box y Jenkins incluyendo modelos de la función de transferencia, modelos de intervención y modelos en retardos distribuidos, es la siguiente:

PROC ARIMA opciones;
BY variables;
IDENTIFY VAR=variable opciones;
ESTIMATE opciones;
FORECAST opciones;

Las opciones de PROC ARIMA son DATA=conjunto de datos de entrada y OUT = conjunto de datos de salida con predicciones.

La sentencia IDENTIFY especifica la variable que contiene los datos a modelizar, las series en diferencias a utilizar si es necesario y halla estadísticos para ayudar a identificar el modelo de ajuste. Es conveniente utilizar una sentencia IDENTIFY para cada serie a ajustar. Sus opciones son las siguientes:

ALPHA= nivel de significación	CENTER centra la serie en su media
CLEAR borra todos los modelos antiguos	CROSSCORR= variable (d11, d12, ..., d1k)
DATA= conjunto de datos de la serie	ESFAC halla la función de autocorrelación muestral extendida para la identificación AR y MA
MINIC usa criterios de información para la identificación ARMA	NLAG= número de retardos a considerar
NOMISS no usa valores desaparecidos	NOPRINT no imprime el output
OUTCOV= conjunto de datos de salida con covarianzas, autocorrelaciones y sus inversas, autocorrelaciones parciales y covarianzas cruzadas	P= (p_{min}: p_{max}) opción de ESACAF, MINIC y SCAN que determina el tamaño de la tabla generada para la identificación

PERROR= ($p_{\varepsilon,min}$: $p_{\varepsilon,max}$) opción de ESACAF, MINIC y SCAN que determina el rando de órdenes del modelo AR que estima los errores	Q= (q_{min}: q_{max}) opción de ESACAF, MINIC y SCAN que determina el tamaño de la tabla generada para la identificación
SCAN halla estimaciones de cuadrados de correlaciones canónicas para tentar la identificación de modelos AR MA	STATIONARITY=(DICKEY= órdenes AR DLAG= s) para realizar el contraste de Dickey-Fuller de órdenes (a,b,c,...) y retardo *s*.
STATIONARITY=(ADF= órdenes AR DLAG= s) realiza el contraste de Dickey-Fuller aumentado de órdenes y retardo dados. Los órdenes AR pueden ser (a,b,c,...)	STATIONARITY=(PP= órdenes AR) realiza el contraste de Phillips-Perron de órdenes y retardo dados. Los órdenes AR pueden ser (a,b,c,...)
STATIONARITY=(RW= órdenes AR) realiza el contraste del camino aleatorio	VAR= variable que contiene las series de tiempo a analizar

Las opciones de la sentencia ESTIMATE pueden clasificarse como sigue:

Opciones para la definición del modelo y control de los estadísticos de diagnóstico

ALTPARM especifica la parametrización alternativa de escala completa en el modelo de la función de transferencia	INPUT= variable INPUT= (función_transferencia variable...) especifica las variables de entrada y su función de transferencia. Las variables utilizadas en INPUT= opción deben estar incluidas en CROSSCORR= lista en la instrucción previa IDENTIFY. Si se especifica una diferencia en CROSSCORR= lista, entonces la serie diferenciada utiliza como input la función de transferencia.	METHOD=ML | ULS | CLS especifica el método de estimación (máxima verosimilitud y mínimos cuadrados incondicionales y condicionales
NOCONSTANT para estimar modelo sin constante	NODF para que en el divisor de la estimación de la varianza no haya grados de libertad	NOPRINT suprime el output
P= orden de la parte AR del modelo	PLOT gráfica las funciones de autocorrelación de los residuos	Q= orden de la parte MA del modelo

Opciones para almacenar resultados en conjuntos de datos SAS

OUTEST= conjunto de datos SAS que contiene los parámetros estimados y las correlacioneso covarianzas	OUTCORR= conjunto de datos SAS que contiene las correlaciones de los parámetros estimados	OUTCOV= conjunto de datos SAS que contiene las covarianzas de los parámetros estimados
OUTMODEL= conjunto de datos en el que se escribe el modelo y parámetros estimados	OUTSTAT=conjunto de datos en el que se escriben los estadísticos de diagnóstico del modelo	

Opciones para especificar valores de parámetros

AR= lista de valores iniciales para los parámetros AR del modelo	INITVAL= valores iniciales para los parámetros de la función de transferencia del modelo	MA= lista de valortes iniciales para los parámetros MA del modelo
U =valor para el parámetro constante	NOEST no realiza estimación iterativa y usa para el modelo los valores de AR, MA , MU e INITVAL dados	

Opciones para controlar el proceso iterativo de estimación

BACKLIM= -n omite *n* residuos iniciales de la suma de cuadrados de la función de verosimilitud	CONVERGE= valor máximo de distancia entre dos iteraciones sucesivas para que se produzca la convergencia	DELTA= valor de la perturbación para hallar derivadas numéricas
GRID imprime la suma de cuadrados del error en una pequeña malla alrededor de las estimaciones	GRIDVAL= n para controlar el espaciado de la malla creada por la opción GRID	MAXITER= número máximo de iteraciones para la convergencia
NOLS para usar mínimos cuadrados incondicionales	NOSTABLE permite estimar AR y MA sin estacionariedad ni invertibilidad	PRINTALL imprime toda la historia de las iteraciones
SINGULAR= valor para chequear la singularidad		

Las opciones de la sentencia FORECAST son las siguientes:

ALIGN= BEGINNING| BEG| B, MIDDLE|MID|M, o ENDING|END|E opciones para controlar la alineación de fechas	ALPHA= n sitúa el tamaño de los límites de confianza de las predicciones	BACK= número de observaciones para predicción relativas a los últimos valores reales de la serie
ID= variable que identifica períodos de tiempo asociados con observaciones	INTERVAL= intervalo de tiempo entre las observaciones	LEAD= número de predicciones a obtener a partir del último valor real de la serie
NOOUTALL sólo incluye la última predicción en la salida	NOPRINT suprime el output	OUT= conjunto de datos que recoge las predicciones
PRINTALL para impresión larga	SIGSQ=varianza de la fórmula de las predicciones y sus intervalos	

Como primer ejemplo modelizaremos una serie *x* mensual de datos desde el 31 de diciembre de 1948 relativos a pasajeros de líneas aéreas en un aeropuerto contenida en el archivo de datos SAS *seriesg.sas7bdat*. Evidentemente comenzamos graficando la serie de datos originales y su logaritmo y analizando la stacionaridad y la estacionalidad mediante la sintaxis siguiente siguiente:

```
ods graphics on;
data datos;
   set ejemplos.seriesg;
   xlog=log(x);
run;
proc timeseries data=datos plot=(series decomp corr periodogram);
   var x xlog;
   id date interval=month;
   spectra p s / adjmean;
run;
```

En la salida observamos el gráfico de la serie *x* (Figura 7-100), las funciones de autocorrelación y autocrrelación parcial para la serie y su primera diferencia (Figura 7-101), su descomposición en componentes (Figura 7-102) y el periodograma por periodos (Figura 7-103).

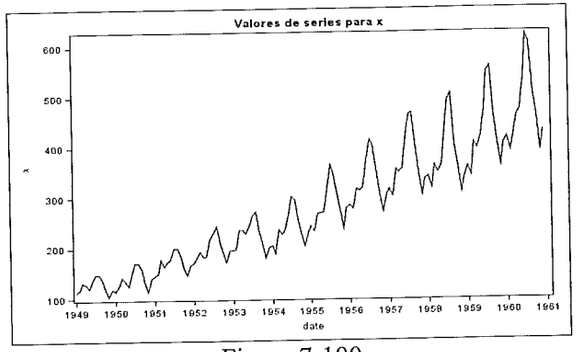


Figura 7-100

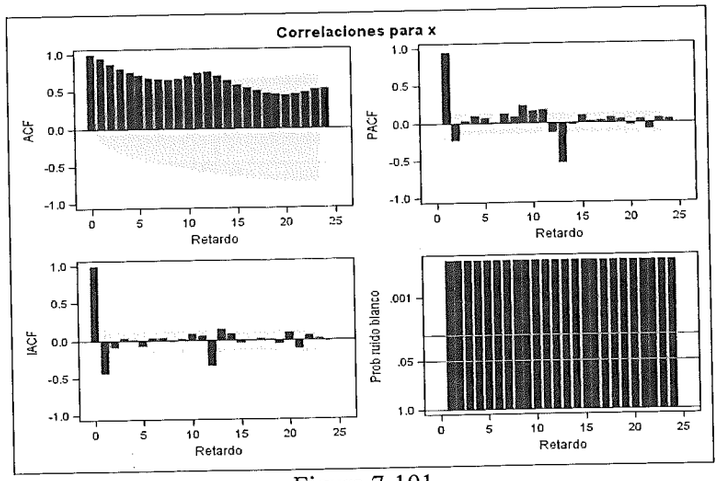


Figura 7-101

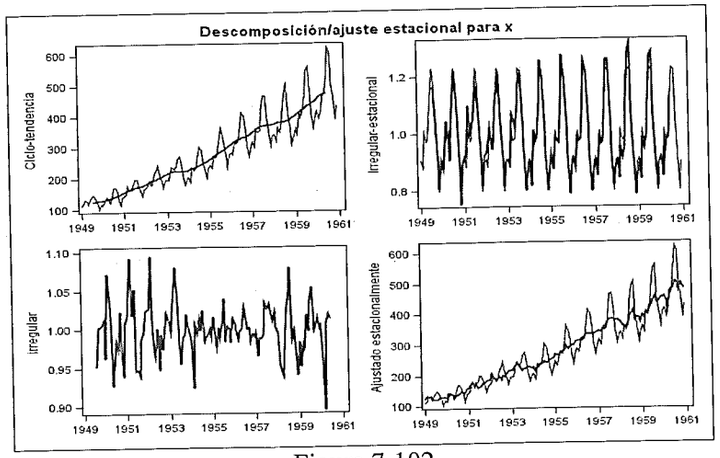


Figura 7-102

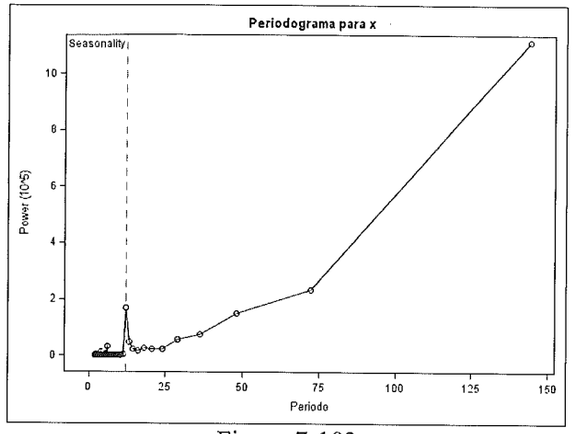


Figura 7-103

Tanto el gráfico de la serie, como su descomposición en componentes y el periodograma muestran que la serie es estacional de periodo 12 (según la abscisa del segundo pico del periodograma). Además, las funciones de autocorrelación y autocorrelación parcial tienen una estructura de abanico con repetición de secuencias ni tienen términos significativos múltiplos del periodo estacional. El gráfico de la serie también muestra claramente que no hay estacionariedad ni en media ni en varianza porque tiene una estructura claramente creciente y poco estable. Por otro lado, la función de autocorrelación de la serie decrece muy lentamente, lo que indica falta de estacionaridad.

La función de autocorrelación de la primera diferencia presenta términos significativos en múltiplos del período estacional, lo que indica que la primera diferencia de la serie no es estacionaria y puede ser conveniente una diferenciación estacional.

Dado el problema de no estacionaridad en varianza de la serie original (tendencia demasiado creciente con mantenimineto creciente de la estacionalidad), es conveniente analizar la serie en logaritmos, tarea que ya se ha incluido en la sintaxis del procedimiento TIMESERIES. En la salida observamos el gráfico de la serie *log*(*x*) (Figura 7-104), las funciones de autocorrelación y autocrrelación parcial para la serie *log*(*x*) y su primera diferencia (Figura 7-105), su descomposición en componentes (Figura 7-106) y el periodograma por periodos (Figura 7-107).

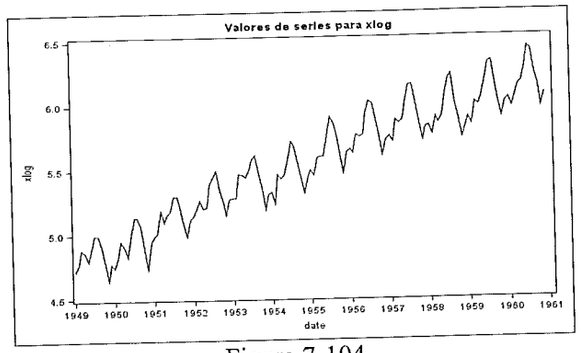


Figura 7-104

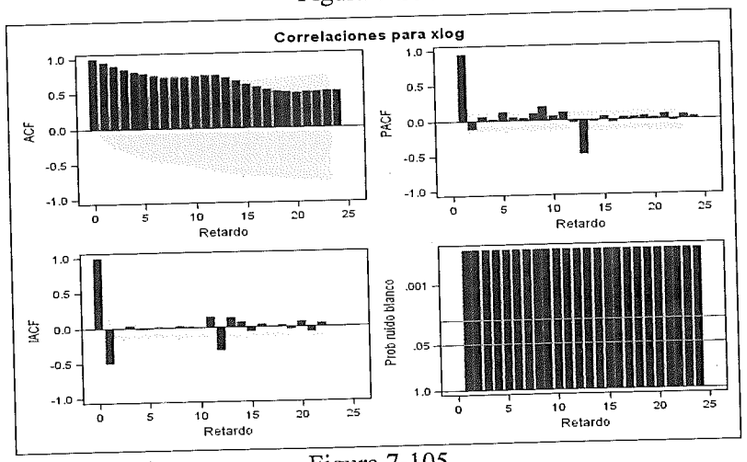


Figura 7-105

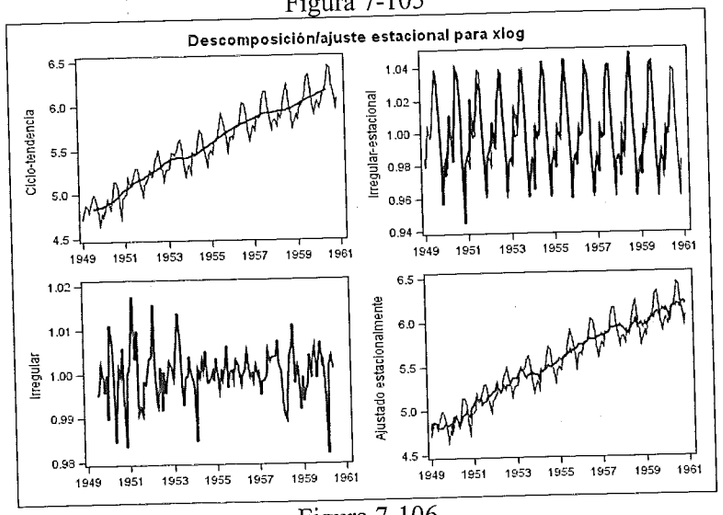


Figura 7-106

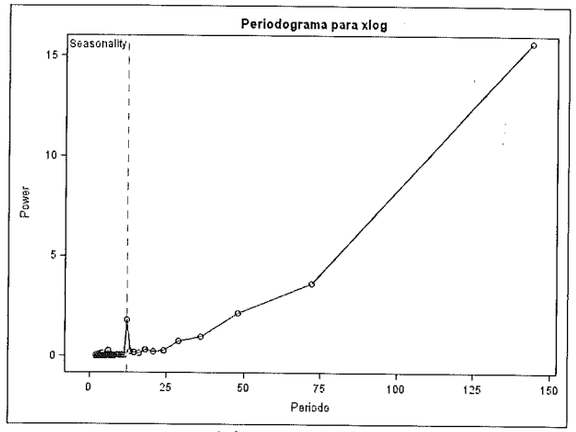


Figura 7-107

Tanto el gráfico de la serie *log(x),* como su descomposición en componentes y el periodograma muestran que la serie es estacional de periodo 12 (según la abscisa del segundo pico del periodograma). Además, las funciones de autocorrelación y autocorrelación parcial tienen una estructura de abanico con repetición de secuencias ni tienen términos significativos múltiplos del periodo estacional. El gráfico de la serie también muestra claramente que no hay estacionariedad ni en media ni en varianza porque tiene una estructura claramente creciente y poco estable. Por otro lado, la función de autocorrelación de la serie decrece muy lentamente, lo que indica falta de estacionaridad.

La función de autocorrelación de la primera diferencia de *log(x)* presenta términos significativos en múltiplos del período estacional, lo que indica que la primera diferencia de la serie no es estacionaria y puede ser conveniente una diferenciación estacional.

Está claro que es conveniente considerar la primera diferencia de la serie en logaritmos tanto para la parte regular como para la parte estacional e intentar la identificación con la siguiente sintaxis SAS basada en el procedimiento ARIMA.

```
ods graphics on;
data datos;
   set ejemplos.seriesg;
   xlog=log(x);
run;
proc arima data=datos;
  identify var=xlog(1,12) nlag=15; /*Una diferencia para la
parte regular de la serie en logaritmos y una diferencia para
la parte estacional con 15 retardos en las funciones de
autocorrelación*/
  run;
```

La Figura 7-108 muestra la salida. Se observa estacionaridad en la serie en logaritmos diferenciada regular y estacionalmente y un patrón de ajuste a un modelo ARIMA(0,0,1)(0,0,1)$_{12}$.

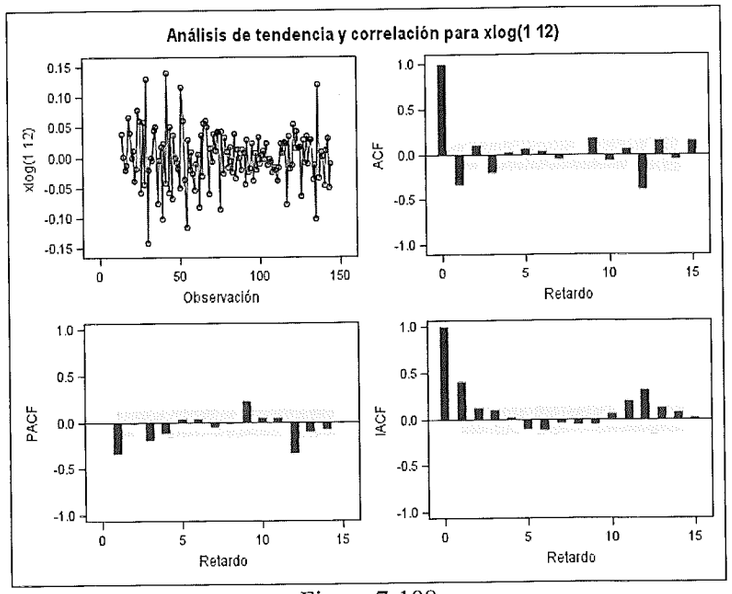


Figura 7-108

Por lo tanto, la serie en logaritmos *log(x)* se ajusta a un modelo ARIMA(0,1,1)×(0,1,1)$_{12}$ que puede ser sin constante.

Una vez identificado el modelo de ajuste de la serie *log(x)*, procedemos a su estimación y cálculo de predicciones con la siguiente sintaxis SAS basada en el procedimiento ARIMA.

```
ods graphics on;
data datos;
   set ejemplos.seriesg;
   xlog=log(x);
run;
proc arima data=datos;
  identify var=xlog(1,12);
  estimate q=(1)(12) noconstant method=uls; /*Se estima un
  modelo ARIMA(0,1,1)x(0,1,1)12 sin constante*/
  forecast out=b lead=24 id=date interval=month printall;
  run; /*Se hacen 24 predicciones siendo date la variable fecha*/
```

La salida es la siguiente:

Proc ARIMA

Nombre de la variable = xlog

Periodo(s) de diferenciación	1,12
Media de series de trabajo	0.000291
Desviación estándar	0.045673
Número de observaciones	131
Observación(es) eliminadas por diferenciación	13

Comprobación de autocorrelación del ruido blanco

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	23.27	6	0.0007	-0.341	0.105	-0.202	0.021	0.056	0.031
12	51.47	12	<.0001	-0.056	-0.001	0.176	-0.076	0.064	-0.387
18	62.44	18	<.0001	0.152	-0.058	0.150	-0.139	0.070	0.016
24	74.27	24	<.0001	-0.011	-0.117	0.039	-0.091	0.223	-0.018

Estimación por mínimos cuadrados no condicional

Parámetro	Estimador	Error estándar	Valor t	Approx Pr > |t|	Retardo
MA1,1	0.39594	0.08149	4.86	<.0001	1
MA2,1	0.61331	0.07961	7.70	<.0001	12

Varianza Estimación	0.001363
Error std Estimación	0.036921
AIC	-484.755
SBC	-479.005
Número de residuales	131

Correlaciones de los estimadores de parámetro

Parámetro	MA1,1	MA2,1
MA1,1	1.000	-0.055
MA2,1	-0.055	1.000

Comprobación de los residuales de autocorrelación

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	5.56	4	0.2349	0.022	0.024	-0.125	-0.129	0.057	0.065
12	8.49	10	0.5816	-0.065	-0.042	0.102	-0.060	0.023	0.007
18	13.23	16	0.6560	0.022	0.039	0.045	-0.162	0.035	0.001
24	24.99	22	0.2978	-0.106	-0.104	-0.037	-0.027	0.219	0.040

Modelo para la variable xlog

Periodo(s) de diferenciación 1,12

No hay ningún término medio en este modelo.

Factores de la media móvil

Factor 1: 1 - 0.39594 B**(1)
Factor 2: 1 - 0.61331 B**(12)

Predicciones para la variable xlog

Obs	Predicción	Error std.	95% Límites de confianza		Real	Residual
142	6.1049	0.0369	6.0325	6.1772	6.1334	0.0285
143	5.9933	0.0369	5.9209	6.0656	5.9661	-0.0271
144	6.0846	0.0369	6.0122	6.1569	6.0684	-0.0161
145	6.1099	0.0369	6.0375	6.1823	.	.
146	6.0558	0.0431	5.9712	6.1403	.	.
147	6.1781	0.0486	6.0830	6.2733	.	.
148	6.1990	0.0534	6.0943	6.3037	.	.
149	6.2312	0.0579	6.1177	6.3447	.	.
150	6.3689	0.0620	6.2473	6.4905	.	.
151	6.5047	0.0659	6.3755	6.6340	.	.
152	6.5013	0.0696	6.3649	6.6377	.	.
153	6.3258	0.0731	6.1826	6.4691	.	.
154	6.2081	0.0764	6.0583	6.3578	.	.
155	6.0644	0.0796	5.9084	6.2204	.	.
156	6.1700	0.0827	6.0079	6.3320	.	.
157	6.2075	0.0904	6.0303	6.3847	.	.
158	6.1534	0.0955	5.9661	6.3407	.	.
159	6.2758	0.1004	6.0789	6.4726	.	.
160	6.2966	0.1051	6.0907	6.5026	.	.
161	6.3288	0.1095	6.1141	6.5435	.	.
162	6.4665	0.1138	6.2435	6.6896	.	.
163	6.6024	0.1179	6.3712	6.8335	.	.
164	6.5989	0.1219	6.3600	6.8379	.	.
165	6.4235	0.1258	6.1769	6.6700	.	.
166	6.3057	0.1295	6.0518	6.5596	.	.
167	6.1620	0.1332	5.9010	6.4230	.	.
168	6.2676	0.1367	5.9996	6.5356	.	.

Las fases de identificación y estimación muestran que la serie en logaritmos se ajusta a la estructura $(1\text{-}B)(1\text{-}B^{12})\log(X_t) = (1\text{-}0{,}39594B)(1\text{-}0{,}61331B^{12})a_t$.

En cuanto a la diagnosis, en la salida vemos que los p-valores de los parámetros estimados son muy pequeños, lo que hace que las estimaciones sean muy significativas. Además, los p-valores del chequeo de autocorrelación residual son mayores que 0,05. Además, la salida muestra los gráficos de normalidad residual (Figura 7-109) que indican aceptación de la normalidad en los residuos. La Figura 7-110 muestra el gráfico de las predicciones.

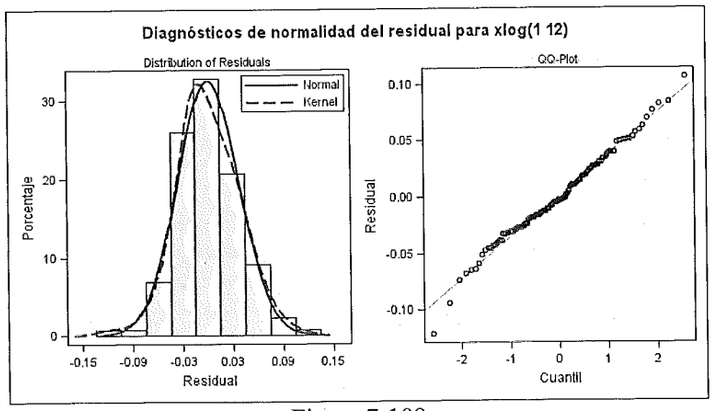


Figura 7-109

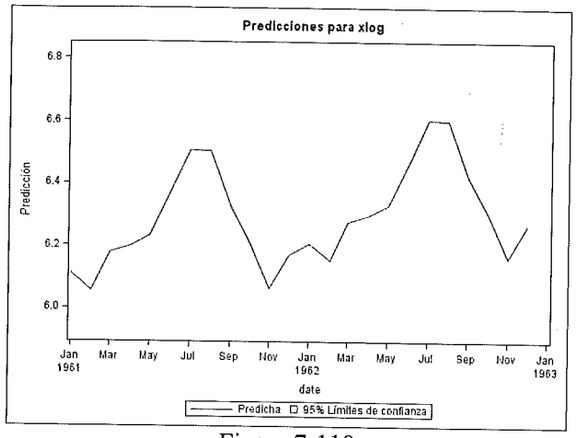


Figura 7-110

No olvidemos que las predicciones P_i se han obtenido para la serie en logaritmos y posteriormente se transformarán en términos de la serie original, es decir, su valor será e^{Pi}.

SAS Y LOS MODELOS ARIMA DE INTERVENCIÓN Y DE FUNCIÓN DE TRANSFERENCIA

La sintaxis del procedimiento ARIMA, que ajusta modelos de series temporales según la metodología de Box y Jenkins incluyendo modelos de la función de transferencia, modelos de intervención y modelos en retardos distribuidos, es la siguiente:

PROC ARIMA opciones;
BY variables;
IDENTIFY VAR=variable opciones;
ESTIMATE opciones;
OUTLIER opciones;
FORECAST opciones;

Las opciones de PROC ARIMA son DATA=conjunto de datos de entrada y OUT=conjunto de datos de salida con predicciones.

La sentencia IDENTIFY especifica la variable que contiene los datos a modelizar, las series en diferencias a utilizar si es necesario y halla estadísticos para ayudar a identificar el modelo de ajuste. Es conveniente utilizar una sentencia IDENTIFY para cada serie a ajustar. Sus opciones más habituales son las siguientes:

Identify var = varible	Identifica la variable
Identify var = variable(1)	Identifica la primera diferencia regular de la variable

Identify var = variable(1,1)	Identifica la segunda diferencia regular de la variable
Identify var = variable(2)	Identifica variable(t) – variable(t-2)
Identify var = variable(12)	Identifica la primera diferencia estacional
Identify var = variable(1,12)	Identifica una diferencia regular y una estacional

identify var=y crosscorr=(x1x2)
Identifica la variable y objeto de predicción y las variables auxiliares de transferencia $x1$ y $x2$.

identify var=y(1) crosscorr=(x1(1) x2(1))
Identifica la primera diferencia de y y las primeras diferencias de las variables auxiliares de transferencia $x1$ y $x2$.

La sentencia ESTIMATE especifica los criterios de estimación. Sus opciones más habituales son las siguientes:

Estimate p = 4	Estima un modelo AR(4)
Estimate p = (1,4)	Estima un modelo AR con retardos 1 y 4 (los demás nulos)
Estimate p = (1 12 13)	Estima un modelo AR con retardos 1 y 12 y 13
Estimate p = (1)(12)	Estima un retardo AR en la parte regular y otro en la estacional

Los esquemas AR según el operador diferencia B serían los siguientes:

P=4	$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_4 B^4)$
P=(1 4)	$(1-\phi_1 B-\phi_4 B^4)$

P=(1 12 13)	$(1-\phi_1 B-\phi_{12} B^{12}-\phi_{13} B^{13})$
P=(1)(12)	$(1-\phi_1 B)(1-\phi_{12} B^{12})$

P=(1 12 13)	$(1-\phi_1 B-\phi_{12} B^{12}-\phi_{13} B^{13})$
P=(1)(12)	$(1-\phi_1 B)(1-\phi_{12} B^{12})$

De forma similar se definen los esquemas de medias móviles MA:

Estimate q = 4	Estima un modelo MA(4)
Estimate q = (1,4)	Estima un modelo MA con retardos 1 y 4 (los demás nulos)
Estimate q = (1 12 13)	Estima un modelo MA con retardos 1 y 12 y 13
Estimate q = (1)(12)	Estima un retardo MA en la parte regular y otro en la estacional

Hay que tener presente que los esquemas anteriores son combinables.

La sentencia OUTLIERS identifica valores atípicos, escalones e impulsos.

Como primer ejemplo consideramos la serie *ventas* con datos mensuales desde enero de 1988 hasta abril de 1996 contenida en el fichero SAS de nombre *test.sas7bdat* (la variable *date* contiene los meses y la variable *i* el índice temporal) e intentamos ajustarla al modelo ARIMA más adecuado con análisis de la intervención para realizar 6 predicciones.

En el capítulo 3 habíamos identificado esta serie como ARIMA(1,1,1). En esta ocasión estimaremos el modelo con análisis de la intervención y calcularemos seis predicciones. La sintaxis SAS adecuada sería la siguiente:

```
data datos;
   set ejemplos.test;
run;

proc arima data=ejemplos.test;
   identify var=ventas(1) nlag=24;
   estimate p=1 q=1;
   outlier;
   forecast lead=6 interval=month id=date out=resultados;
   run;
```

En primer lugar, obtenemos información relativa a la identificación. El p-valor muy bajo de la chi-cuadrado (<0,0001) indica que la serie inicial no es un ruido blanco (de lo contrario, no se podría modelizar). El panel de correlaciones de la Figura 7-111 muestra que la serie diferenciada puede ajustarse a un modelo ARMA(1,1).

Proc ARIMA

Nombre de la variable = VENTAS

Periodo(s) de diferenciación	1
Media de series de trabajo	0.660589
Desviación estándar	2.011543
Número de observaciones	99
Observación(es) eliminadas por diferenciación	1

Comprobación de autocorrelación del ruido blanco

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	154.44	6	<.0001	0.828	0.591	0.454	0.369	0.281	0.198
12	173.66	12	<.0001	0.151	0.081	-0.039	-0.141	-0.210	-0.274
18	209.64	18	<.0001	-0.305	-0.271	-0.218	-0.183	-0.174	-0.161
24	218.04	24	<.0001	-0.144	-0.141	-0.125	-0.085	-0.040	-0.032

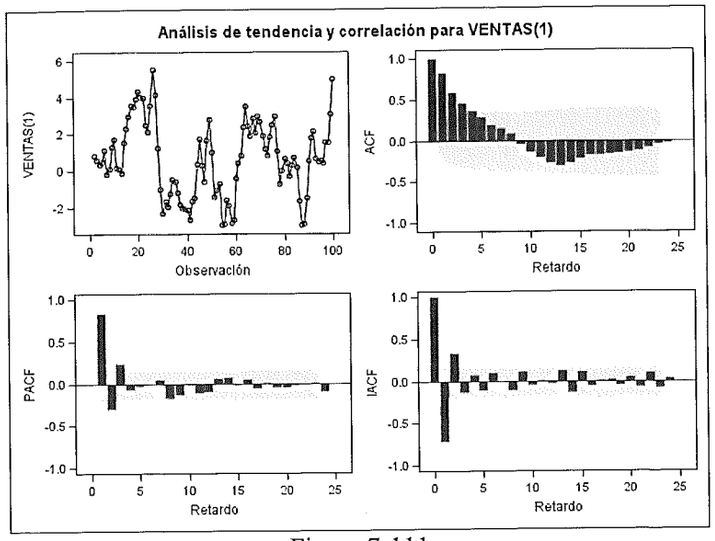


Figura 7-111

A continuación, se presenta la estimación de los parámetros del modelo y de las variables de intervención. Se observa la alta significatividad de los parámetros del modelo (p-valores muy pequeños). Las correlaciones entre los parámetros estimados son pequeñas y los residuos resultan aleatorios porque los p-valores del contraste de la chi-cuadrado son muy altos. Los paneles de correlaciones residuales (Figura 7-112) y normalidad residual (Figura 7-113) constatan la superación de la diagnosis residual.

Estimación por mínimos cuadrados condicional

Parámetro	Estimador	Error estándar	Valor t	Approx Pr > |t|	Retardo
MU	0.89288	0.49391	1.81	0.0738	0
MA1,1	-0.58935	0.08988	-6.56	<.0001	1
AR1,1	0.74755	0.07785	9.60	<.0001	1

Constante Estimación	0.225409
Varianza Estimación	0.904034
Error std Estimación	0.950807
AIC	273.9155
SBC	281.7009
Número de residuales	99

* AIC y SBC no incluyen determinante de la log.

Correlaciones de los estimadores de parámetro

Parámetro	MU	MA1,1	AR1,1
MU	1.000	0.030	0.107
MA1,1	0.030	1.000	0.395
AR1,1	0.107	0.395	1.000

Comprobación de los residuales de autocorrelación

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	3.95	4	0.4127	0.016	-0.044	-0.068	0.145	0.024	-0.094
12	7.03	10	0.7227	0.088	0.087	-0.037	-0.075	0.051	-0.053
18	15.41	16	0.4951	-0.221	-0.033	-0.092	0.086	-0.074	-0.005
24	16.96	22	0.7657	0.011	-0.066	-0.022	-0.032	0.062	-0.047

Modelo para la variable VENTAS

Media estimada	0.892875
Periodo(s) de diferenciación	1

Factores autorregresivos

Factor 1: 1 - 0.74755 B**(1)

Factores de la media móvil

Factor 1: 1 + 0.58935 B**(1)

Resumen de detección de valores extremos

Número máximo buscado	2
Número encontrado	2
Significación empleada	0.05

Se constata que hay dos outliers de tipo aditivo en las observaciones 10 y 67.

Detalles de valores extremos

Obs	Tipo	Estimador	Chi-cuadrado	Approx Prob> ChiSq
10	Aditivo	0.56879	4.20	0.0403
67	Aditivo	0.55698	4.42	0.0355

La ecuación del modelo estimado será la siguiente:

$$\boxed{\begin{array}{l}(1-0{,}74755B)(1-B)VENTAS = 0.892875(1+0.58935B)a_t - \\ -0.56879I_t^{1988} - 0.55698I_t^{1993} \quad I_t^{t0} = 1\ si\ t = t_0 \quad I_t^{t0} = 0\ si\ t \neq t_0\end{array}}$$

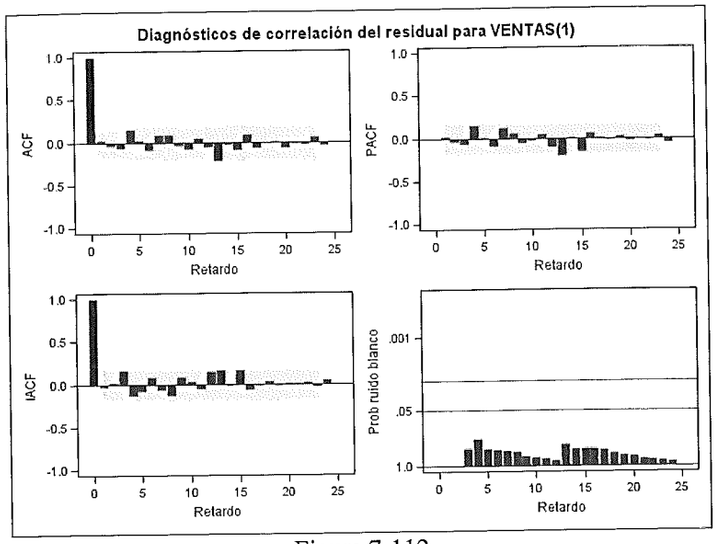


Figura 7-112

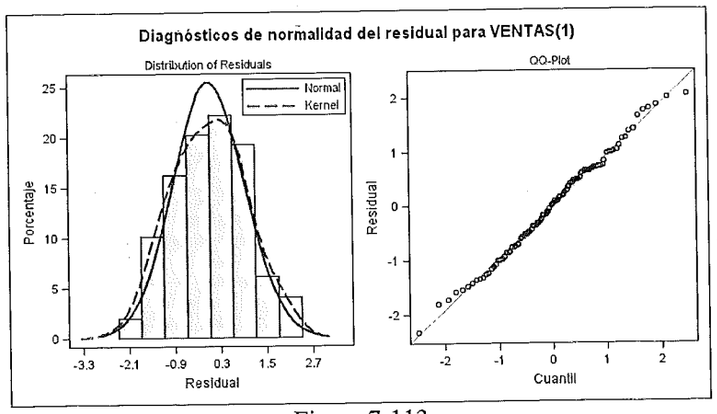


Figura 7-113

Por último se presentan las predicciones pedidas.

Predicciones para la variable VENTAS

Obs	Predicción	Error std.	95% Límites de confianza	
101	171.0320	0.9508	169.1684	172.8955
102	174.7534	2.4168	170.0165	179.4903
103	177.7608	3.9879	169.9445	185.5770
104	180.2343	5.5658	169.3256	191.1430
105	182.3088	7.1033	168.3866	196.2310
106	184.0850	8.5789	167.2707	200.8993

La Figura 7-114 muestra las predicciones de ventas.

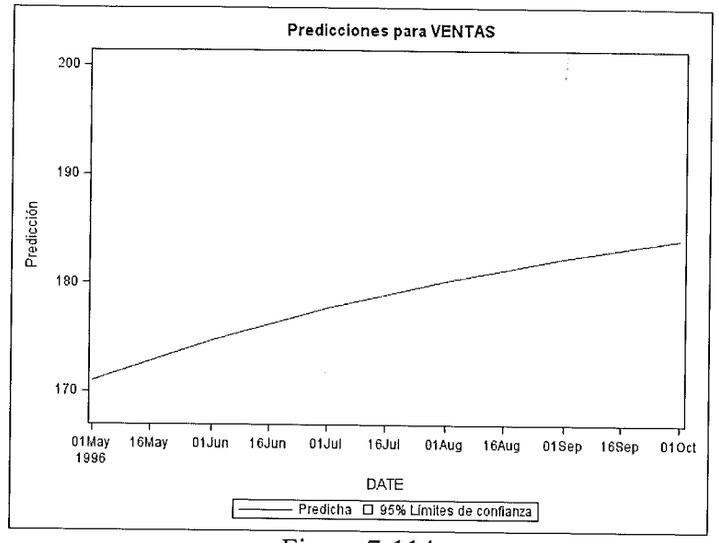


Figura 7-114

Como segundo ejemplo, estimaremos un modelo de la función de transferencia que busca predecir la serie *ventas* con información adicional de la nueva variable auxiliar *precio* que entra en el modelo para influir en las predicciones sobre la variable en estudio. Los datos son mensuales desde enero de 1988 hasta abril de 1996 y están contenidos en el fichero SAS de nombre *mercado.sas7bdat* (la variable *date* contiene los meses y la variable *i* el índice temporal). Se observa la inclusión de la opción *CROSSCORR = variable auxiliar* en la sentencia *identify*. También es necesario incluir la opción *INPUT = variable auxiliar* en la sentencia *estimate*. La sintaxis SAS adecuada sería la siguiente:

```
ods graphics on;
data datos;
   set ejemplos.mercado;
run;
proc arima data=datos;
   identify var=ventas crosscorr=precio;
   estimate input=precio;
run;
```

En la salida obtenemos la estimación de un modelo de la forma:

$$Y_t = \mu + \omega_0 X_t + a_t$$

Obtenemos los siguientes resultados para la estimación del modelo:

Estimación por mínimos cuadrados condicional

Parámetro	Estimador	Error estándar	Valor t	Approx Pr > |t|	Retardo	Variable	Desplazamiento
MU	-199.83602	2.99463	-66.73	<.0001	0	ventas	0
NUM1	9.99299	0.02885	346.38	<.0001	0	precio	0

Constante Estimación	-199.836
Varianza Estimación	69.20761
Error std Estimación	8.319111
AIC	709.4785
SBC	714.6889
Número de residuales	100

* AIC y SBC no incluyen determinante de la log.

El modelo estimado tendrá la ecuación siguiente:

$$\boxed{VENTAS = -199{,}836 + 9{,}99299\,PRECIO + a_t}$$

La figura 7-115 muestra el panel de normalidad residual.

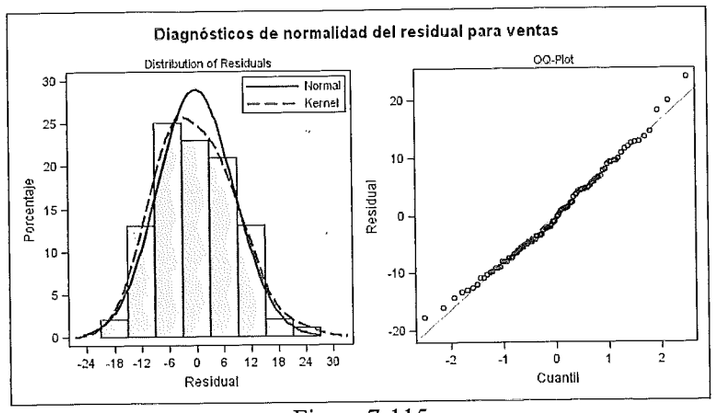


Figura 7-115

Como tercer ejemplo, estimaremos el modelo de la función de transferencia anterior diferenciando una vez todas las variables. Se observa que para diferenciar una vez la variable auxiliar es necesario incluir la opción *INPUT = (1 $ variable auxiliar)* en la sentencia *estimate*. En la sentencia *identify* se introduce la diferencia de las variables con la notación habitual. La sintaxis SAS adecuada sería la siguiente:

```
ods graphics on;
data datos;
   set ejemplos.mercado;
run;
```



```
proc arima data=datos;
   identify var=ventas(1) crosscorr=precio(1);
   estimate input=(1 $ precio);
run;
```

En la salida obtenemos la estimación de un modelo de la forma:

$$(1-B)Y_t = \mu + \omega_0(1-B)X_{t-1} + a_t$$

Obtenemos los siguientes resultados para la estimación del modelo:

Estimación por mínimos cuadrados condicional

Parámetro	Estimador	Error estándar	Valor t	Approx Pr > |t|	Retardo	Variable	Desplazamiento
MU	14.13693	4.66993	3.03	0.0032	0	ventas	0
NUM1	-4.18059	0.93358	-4.48	<.0001	0	precio	1

Constante Estimación	14.13693
Varianza Estimación	2055.822
Error std Estimación	45.34117
AIC	1027.677
SBC	1032.847
Número de residuales	98

* AIC y SBC no incluyen determinante de la log.

El modelo estimado tiene la ecuación siguiente:

$$\boxed{(1-B)VENTAS_t = 14{,}13693 - 4{,}18059(1-B)PRECIO_{t-1} + a_t}$$

La figura 7-116 muestra el panel de normalidad residual.

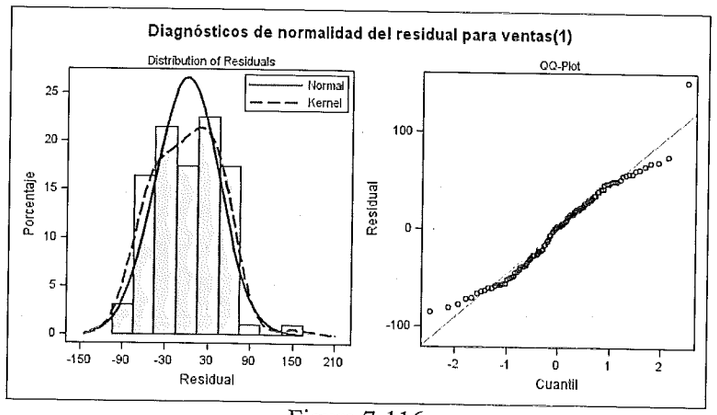


Figura 7-116

Como cuarto ejemplo, estimaremos el modelo del primer ejemplo añadiendo como nueva variable auxiliar la variable *ingreso*. Se trata por tanto de ajustar la evolución de la variable ventas teniendo en cuenta los efectos sobre ella de las variables *precio* e *ingreso*. La sintaxis SAS sería la siguiente:

En la salida obtenemos la estimación de un modelo de la forma:

$$Y_t = \mu + \omega_1 X_{1,t} + \omega_2 X_{2,t} + a_t$$

Obtenemos los siguientes resultados para la estimación del modelo:

Estimación por mínimos cuadrados condicional

Parámetro	Estimador	Error estándar	Valor t	Approx Pr > |t|	Retardo	Variable	Desplazamiento
MU	-203.58109	10.39962	-19.58	<.0001	0	ventas	0
NUM1	10.06727	0.19957	50.44	<.0001	0	precio	0
NUM2	-0.0000417	0.0001109	-0.38	0.7076	0	ingreso	0

Constante Estimación	-203.581
Varianza Estimación	69.81923
Error std Estimación	8.35579
AIC	711.3327
SBC	719.1482
Número de residuales	100

* AIC y SBC no incluyen determinante de la log.

El modelo estimado tendrá la ecuación siguiente:

$$\boxed{VENTAS = -202{,}581 + 10{,}06727\,PRECIO - 0{,}000417\,INGRESO + a_t}$$

Observamos que le p-valor de la variable ingreso es muy alto, lo que indica que esta variable no es significativa en el modelo. Además, su parámetro estimado es casi nulo.

La figura 7-117 muestra el panel de normalidad residual.

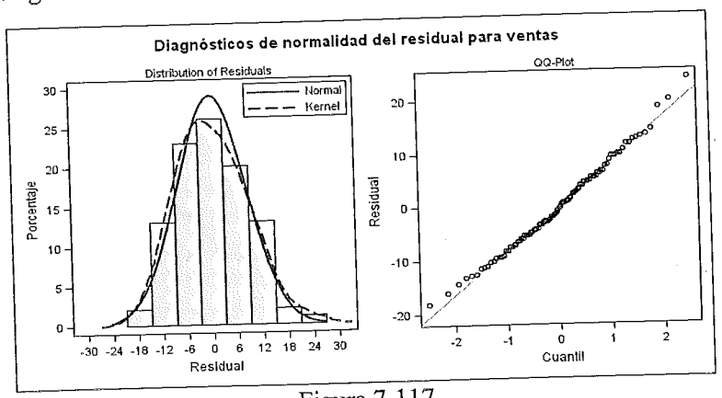


Figura 7-117

Como quinto ejemplo, se trata por tanto de ajustar la evolución de la variable *ventas* teniendo en cuenta los efectos sobre ella de las variables *precio* e *ingreso* (el mismo modelo del ejemplo anterior), pero teniendo en cuenta ahora una estructura ARMA(1,1) para los residuos. La sintaxis SAS sería la siguiente:

```
ods graphics on;
data datos;
   set ejemplos.mercado;
run;
proc arima data=datos;
   identify var=ventas crosscorr=(precio ingreso);
   estimate p=1 q=1 input=(precio ingreso);
run;
```

En la salida obtenemos la estimación de un modelo de la forma:

$$Y_t = \mu + \omega_1 X_{1,t} + \omega_2 X_{2,t} + \frac{(1-\theta_1 B)}{(1-\phi_1 B)} a_t$$

Obtenemos los siguientes resultados para la estimación del modelo:

Estimación por mínimos cuadrados condicional

Parámetro	Estimador	Error estándar	Valor t	Approx Pr > |t|	Retardo	Variable	Desplazamiento
MU	-112.13151	26.86675	-4.17	<.0001	0	ventas	0
MA1,1	0.68951	0.08561	8.05	<.0001	1	ventas	0
AR1,1	0.99999	0.01108	90.29	<.0001	1	ventas	0
NUM1	7.98862	0.59857	13.35	<.0001	0	precio	0
NUM2	0.0009477	0.0003312	2.86	0.0052	0	ingreso	0

Constante Estimación	-1.45E-6
Varianza Estimación	74.57947
Error std Estimación	8.635941
AIC	719.8449
SBC	732.8708
Número de residuales	100

* AIC y SBC no incluyen determinante de la log.

Correlaciones de los estimadores de parámetro

Parámetro	de la variable	ventas MU	ventas MA1,1	ventas AR1,1	precio NUM1	ingreso NUM2
ventas	MU	1.000	-0.303	-0.023	-0.970	0.882
ventas	MA1,1	-0.303	1.000	0.277	0.323	-0.358
ventas	AR1,1	-0.023	0.277	1.000	0.069	-0.278
precio	NUM1	-0.970	0.323	0.069	1.000	-0.936
ingreso	NUM2	0.882	-0.358	-0.278	-0.936	1.000

Comprobación de los residuales de autocorrelación

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	9.80	4	0.0439	-0.118	0.162	-0.209	0.002	0.093	-0.029
12	25.35	10	0.0047	0.003	0.108	0.302	-0.048	0.065	-0.169
18	27.35	16	0.0378	0.033	0.082	0.036	-0.001	0.019	0.084
24	33.57	22	0.0542	0.036	0.089	-0.152	0.016	-0.037	0.117

Modelo para la variable ventas

Término independiente estimado -112.132

Factores autorregresivos

Factor 1: 1 - B**(1)

Factores de la media móvil

Factor 1: 1 - 0.68951 B**(1)

Número de entrada 1

Variable de entrada	precio
Factor de regresión global	7.988625

Número de entrada 2

Variable de entrada	ingreso
Factor de regresión global	0.000948

El modelo estimado tendrá la ecuación siguiente:

$$\boxed{VENTAS = -112{,}13151 + 7{,}98862\,PRECIO + 0{,}0009977\,INGRESO + \frac{(1-0{,}68951B)}{1-B}a_t}$$

Observamos que los p-valores de los parámetros estimados son muy bajos, lo que indica buena significatividad.

La figura 7-118 muestra el panel de normalidad residual.

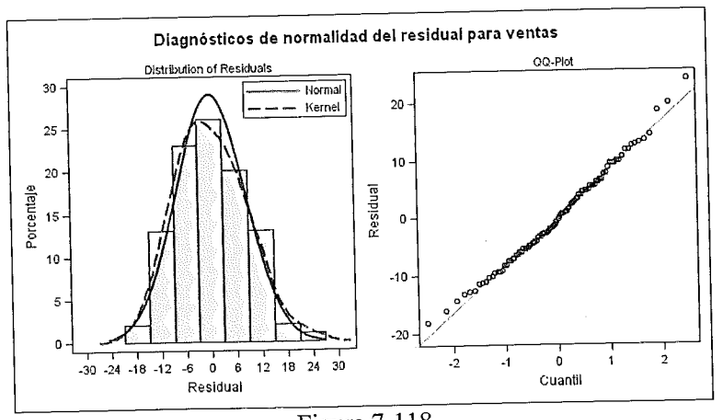


Figura 7-118

Como sexto ejemplo, consideramos un modelo para predecir las ventas en presencia de un impulso que acontece, fruto de un esfuerzo especial en las ventas en el mes de marzo de 1992. Construiremos la intervención como un indicador y asumimos un modelo ARMA(1,1) para las ventas. La sintaxis SAS sería la siguiente:

```
data a;
      set ejemplos.mercado;
      ad = (date = '1mar1992'd);
   run;
proc arima data=a;
      identify var=ventas crosscorr=ad;
      estimate p=1 q=1 input=ad;
   run;
```

La salida es la siguiente:

Proc ARIMA

Nombre de la variable = ventas

Media de series de trabajo	796.6084
Desviación estándar	288.273
Número de observaciones	100

Comprobación de autocorrelación del ruido blanco

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	495.11	6	<.0001	0.956	0.926	0.898	0.870	0.843	0.804
12	829.75	12	<.0001	0.774	0.748	0.717	0.687	0.659	0.630
18	1035.66	18	<.0001	0.603	0.575	0.547	0.518	0.491	0.457
24	1142.62	24	<.0001	0.435	0.408	0.383	0.354	0.327	0.305

Correlación de ventas y ad

Varianza de la entrada =	0.0099
Número de observaciones	100

WARNING: The model defined by the new estimates is unstable. The iteration process has been terminated.

WARNING: Puede que los estimadores no hayan convergido

Resumen de la optimización de la estimación ARIMA

Método de estimación	Cuadrados mínimos condicionales
Parámetros estimados	4
Criterios de terminación	Cambio relativo máximo en los estimadores
Valor de terminación de la iteración	0.001
Valor de los criterios	1.771879
Valor absoluto máximo del gradiente	674148
Cambio de R-cuadrado desde la última iteración	0.41442
Función objetiva	Suma de los residuales cuadrados
Valor de la función objetiva	215009.4
Coeficiente Lambda de Marquardt	1E-6
Valor de perturbación para el cálculo de derivadas por métodos numéricos	0.001
Iteraciones	17
Mensaje de warning	Puede que los estimadores no hayan convergido

Estimación por mínimos cuadrados condicional

Parámetro	Estimador	Error estándar	Valor t	Approx Pr > |t|	Retardo	Variable	Desplazamiento
MU	391.96039	45.42155	8.63	<.0001	0	ventas	0
MA1,1	0.28701	0.11379	2.52	0.0133	1	ventas	0
AR1,1	1.00000	0.0074957	133.41	<.0001	1	ventas	0
NUM1	6.79937	38.41687	0.18	0.8599	0	ad	0

Constante Estimación	0.000026
Varianza Estimación	2239.681
Error std Estimación	47.32527
AIC	1059.114
SBC	1069.535
Número de residuales	100

* AIC y SBC no incluyen determinante de la log.

Correlaciones de los estimadores de parámetro

Parámetro de la variable		ventas MU	ventas MA1,1	ventas AR1,1	ad NUM1
ventas	MU	1.000	-0.062	-0.018	-0.009
ventas	MA1,1	-0.062	1.000	0.371	0.153
ventas	AR1,1	-0.018	0.371	1.000	0.054
ad	NUM1	-0.009	0.153	0.054	1.000

Comprobación de los residuales de autocorrelación

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	7.80	4	0.0991	-0.186	-0.033	0.046	0.096	0.158	-0.047
12	16.10	10	0.0968	-0.022	0.214	0.041	0.062	0.112	0.098
18	21.71	16	0.1526	-0.044	0.105	0.081	-0.005	0.153	0.057
24	27.02	22	0.2104	0.063	0.075	0.054	0.154	-0.066	0.013

Modelo para la variable ventas

Término independiente estimado 391.9604

Factores autorregresivos

Factor 1: 1 - 1 B**(1)

Factores de la media móvil

Factor 1: 1 - 0.28701 B**(1)

Número de entrada 1

Variable de entrada	ad
Factor de regresión global	6.799368

El modelo estimado tiene la ecuación siguiente:

$$\boxed{(1-B)VENTAS_t = 291{,}9604 + (1-0{,}28701)a_t + 6{,}799368 I_t^{to}}$$

$$I_t^{t0} = \begin{cases} 1 \; si \; t_0 = 1 \; de \; Marzo \; 1992 \\ 0 \; en \; otro \; caso \end{cases}$$

Observamos que los p-valores de los parámetros estimados son muy bajos, lo que indica buena significatividad. El parámetro de regresión es poco significativo porque su valor es muy alto (0,8599).

La figura 7-119 muestra el panel de normalidad residual con un buen ajuste de los residuos a la distribución normal.

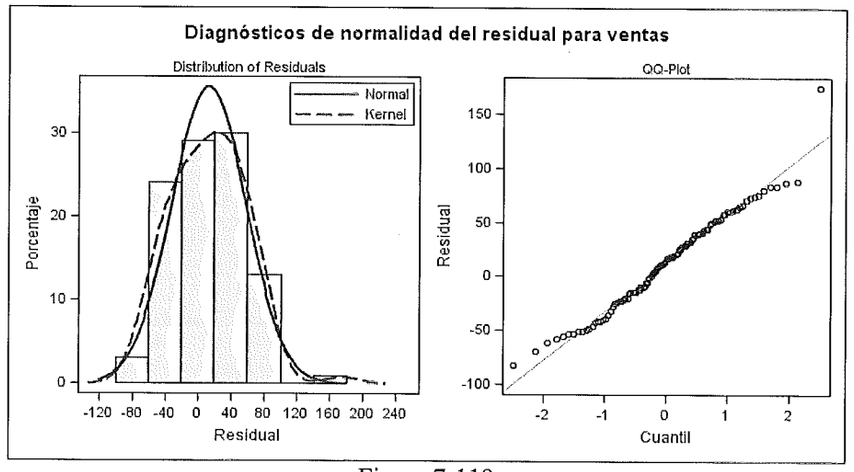


Figura 7-119

Como séptimo ejemplo, consideramos un modelo para predecir las ventas en presencia de un escalón que acontece, fruto de un aumento perenne dela demanda a partir del 1 de julio de 1993. Construiremos la intervención como un indicador y asumimos un modelo ARMA(1,1) para las ventas. La sintaxis SAS sería la siguiente:

```
ods graphics on;
data b;
      set ejemplos.mercado;
      bd = (date >= '1jul1993'd);
   run;

   proc arima data=b;
      identify var=ventas crosscorr=bd;
      estimate p=1 q=1 input=bd;
   run;
```

La salida se presenta a continuación:

Proc ARIMA

Nombre de la variable = ventas

Media de series de trabajo	796.6084
Desviación estándar	288.273
Número de observaciones	100

Comprobación de autocorrelación del ruido blanco

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	495.11	6	<.0001	0.956	0.926	0.898	0.870	0.843	0.804
12	829.75	12	<.0001	0.774	0.748	0.717	0.687	0.659	0.630
18	1035.66	18	<.0001	0.603	0.575	0.547	0.518	0.491	0.457
24	1142.62	24	<.0001	0.435	0.408	0.383	0.354	0.327	0.305

Correlación de ventas y bd

Varianza de la entrada = 0.2244
Número de observaciones 100

WARNING: The model defined by the new estimates is unstable. The iteration process has been terminated.

WARNING: Puede que los estimadores no hayan convergido

Resumen de la optimización de la estimación ARIMA

Método de estimación	Cuadrados mínimos condicionales
Parámetros estimados	4
Criterios de terminación	Cambio relativo máximo en los estimadores
Valor de terminación de la iteración	0.001
Valor de los criterios	1.234863
Valor absoluto máximo del gradiente	1836173
Cambio de R-cuadrado desde la última iteración	0.387317
Función objetiva	Suma de los residuales cuadrados
Valor de la función objetiva	204058.2
Coeficiente Lambda de Marquardt	1E-6
Valor de perturbación para el cálculo de derivadas por métodos numéricos	0.001
Iteraciones	15
Mensaje de warning	Puede que los estimadores no hayan convergido

Estimación por mínimos cuadrados condicional

Parámetro	Estimador	Error estándar	Valor t	Approx Pr > |t|	Retardo	Variable	Desplazamiento
MU	337.39992	39.06050	8.64	<.0001	0	ventas	0
MA1,1	0.53260	0.10312	5.16	<.0001	1	ventas	0
AR1,1	1.00000	0.0055106	181.47	<.0001	1	ventas	0
NUM1	10.20701	40.03884	0.25	0.7993	0	bd	0

Constante Estimación	0.00002
Varianza Estimación	2125.606
Error std Estimación	46.10429
AIC	1053.887
SBC	1064.307
Número de residuales	100

* AIC y SBC no incluyen determinante de la log.

Correlaciones de los estimadores de parámetro

Parámetro de la variable		ventas MU	ventas MA1,1	ventas AR1,1	bd NUM1
ventas	MU	1.000	-0.045	-0.026	0.002
ventas	MA1,1	-0.045	1.000	0.636	-0.061
ventas	AR1,1	-0.026	0.636	1.000	-0.205
bd	NUM1	0.002	-0.061	-0.205	1.000

Comprobación de los residuales de autocorrelación

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	13.82	4	0.0079	0.054	0.084	0.134	0.207	0.233	0.070
12	41.76	10	<.0001	0.103	0.310	0.176	0.166	0.207	0.200
18	61.47	16	<.0001	0.060	0.185	0.193	0.113	0.231	0.145
24	82.09	22	<.0001	0.170	0.175	0.168	0.236	0.039	0.121

Modelo para la variable ventas

Término independiente estimado 337.3999

Factores autorregresivos

Factor 1: 1 - 1 B**(1)

Factores de la media móvil

Factor 1: 1 - 0.5326 B**(1)

Número de entrada 1

Variable de entrada	bd
Factor de regresión global	10.20701

El modelo estimado tiene la ecuación siguiente:

$$\boxed{(1-B)VENTAS_t = 337{,}39 + (1-0{,}5326)a_t + 10{,}20701 I_t^{to}}$$

$$I_t^{t0} = \begin{cases} 1 \; si \; t_0 > 1 \; de \; Julio \; 1993 \\ 0 \; en \; otro \; caso \end{cases}$$

Observamos que los p-valores de los parámetros estimados son muy bajos, lo que indica buena significatividad. El parámetro de regresión es poco significativo porque su valor es muy alto (0,7993).

Como octavo ejemplo, modelizamos las ventas en función en presencia del efecto de los precios mediante una función de transferencia racional con sólo numerador del tipo:

$$Y_t = \mu + (\omega_0 - \omega_1 B - \omega_2 B^2 - \omega_3 B^3)X_t + a_t$$

La sintaxis SAS adecuada sería la siguiente:

```
ods graphics on;
data datos;
      set ejemplos.mercado;
   run;
proc arima data=datos;
      identify var=ventas crosscorr=precio;
      estimate input=( (1 2 3) precio );
   run;
```

La salida se presenta a continuación:

Proc ARIMA

Nombre de la variable = ventas

Media de series de trabajo	796.6084
Desviación estándar	288.273
Número de observaciones	100

Comprobación de autocorrelación del ruido blanco

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	495.11	6	<.0001	0.956	0.926	0.898	0.870	0.843	0.804
12	829.75	12	<.0001	0.774	0.748	0.717	0.687	0.659	0.630
18	1035.66	18	<.0001	0.603	0.575	0.547	0.518	0.491	0.457
24	1142.62	24	<.0001	0.435	0.408	0.383	0.354	0.327	0.305

Correlación de ventas y precio

Varianza de la entrada =	831.5002
Número de observaciones	100

Estimación por mínimos cuadrados condicional

Parámetro	Estimador	Error estándar	Valor t	Approx Pr > |t|	Retardo	Variable	Desplazamiento
MU	-199.63309	3.20428	-62.30	<.0001	0	ventas	0
NUM1	9.62655	0.20030	48.06	<.0001	0	precio	0
NUM1,1	-0.0061459	0.21029	-0.03	0.9767	1	precio	0
NUM1,2	-0.28054	0.21014	-1.34	0.1852	2	precio	0
NUM1,3	-0.08588	0.20053	-0.43	0.6695	3	precio	0

Constante Estimación	-199.633
Varianza Estimación	70.11987
Error std Estimación	8.373761
AIC	692.4106
SBC	705.2842
Número de residuales	97

* AIC y SBC no incluyen determinante de la log.

Correlaciones de los estimadores de parámetro

Parámetro de la variable		ventas MU	precio NUM1	precio NUM1,1	precio NUM1,2	precio NUM1,3
ventas	MU	1.000	-0.153	0.067	0.005	-0.084
precio	NUM1	-0.153	1.000	0.421	0.227	0.313
precio	NUM1,1	0.067	0.421	1.000	-0.375	-0.229
precio	NUM1,2	0.005	0.227	-0.375	1.000	-0.423
precio	NUM1,3	-0.084	0.313	-0.229	-0.423	1.000

Comprobación de los residuales de autocorrelación

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	8.45	6	0.2071	-0.029	0.230	-0.138	-0.081	0.053	-0.030
12	21.05	12	0.0497	-0.026	0.088	0.203	-0.071	0.025	-0.241
18	22.23	18	0.2219	-0.069	-0.017	-0.037	-0.058	-0.005	-0.019
24	30.02	24	0.1841	0.004	-0.028	-0.212	-0.072	-0.094	0.042

Modelo para la variable ventas

Término independiente estimado -199.633

Número de entrada 1

Variable de entrada precio

Factores numeradores

Factor 1: 9.62655 + 0.00615 B**(1) + 0.28054 B**(2) + 0.08588 B**(3)

Observamos que resultan significativos el término de medias móviles y la constante y la diferencia de orden dos de la función de transferencias (p-valores pequeños). Las diferencias de orden uno y tres de la función de transferencia resultan no significativos.

Como noveno ejemplo, modelizamos las ventas en presencia del efecto de los precios mediante una función de transferencia racional con sólo término con diferencias en el denominador del tipo:

$$\frac{\omega_0}{(1-\delta_1 B)} X_t$$

La sintaxis SAS adecuada sería la siguiente:

```
ods graphics on;
data datos;
      set ejemplos.mercado;
   run;
proc arima data=datos;
      identify var=ventas crosscorr=precio;
      estimate input=( / (1) precio);
   run;
```

La salida se presenta a continuación:

Proc ARIMA

Nombre de la variable = ventas

Media de series de trabajo	796.6084
Desviación estándar	288.273
Número de observaciones	100

Comprobación de autocorrelación del ruido blanco

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	495.11	6	<.0001	0.956	0.926	0.898	0.870	0.843	0.804
12	829.75	12	<.0001	0.774	0.748	0.717	0.687	0.659	0.630
18	1035.66	18	<.0001	0.603	0.575	0.547	0.518	0.491	0.457
24	1142.62	24	<.0001	0.435	0.408	0.383	0.354	0.327	0.305

Correlación de ventas y precio
Varianza de la entrada = 831.5002
Número de observaciones 100

Estimación por mínimos cuadrados condicional

Parámetro	Estimador	Error estándar	Valor t	Approx Pr > |t|	Retardo	Variable	Desplazamiento
MU	-200.02628	3.05787	-65.41	<.0001	0	ventas	
NUM1	9.77014	0.17520	55.77	<.0001	0	precio	
DEN1,1	0.02270	0.01753	1.30	0.1983	1	precio	

Constante Estimación	-200.026
Varianza Estimación	69.47267
Error std Estimación	8.335027
AIC	703.7558
SBC	711.5412
Número de residuales	99

* AIC y SBC no incluyen determinante de la log.

Correlaciones de los estimadores de parámetro

Parámetro de la variable		ventas MU	precio NUM1	precio DEN1,1
ventas	MU	1.000	-0.132	-0.026
precio	NUM1	-0.132	1.000	-0.987
precio	DEN1,1	-0.026	-0.987	1.000

Comprobación de los residuales de autocorrelación

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	8.97	6	0.1755	-0.015	0.216	-0.157	-0.051	0.071	-0.084
12	20.46	12	0.0589	-0.024	0.064	0.226	-0.048	-0.008	-0.209
18	22.14	18	0.2260	-0.039	0.003	-0.060	-0.088	-0.029	-0.019
24	28.80	24	0.2279	0.008	-0.011	-0.192	-0.075	-0.087	0.039

Modelo para la variable ventas

Término independiente estimado -200.026

Número de entrada 1

Variable de entrada	precio
Factor de regresión global	9.770143

Factores denominadores

Factor 1: 1 - 0.0227 B**(1)

Los parámetros estimados son todos significativos.

Para un afunción de transferencia racional completa, se ajustaría una ecuación del tipo:

$$\frac{\omega(B)}{\delta(B)}B^{k}X_{t}$$

La sentencia INPUT tendría en este caso la siguiente estructura:

input=(k \$ (ω-*lags*) / (δ-*lags*) x)

El cálculo de las predicciones con estos tipos de modelos se realiza a través de la sentencia FORECAST. El siguiente ejemplo realiza las fases de identificación, estimación y diagnosis para obtener las predicciones de ventas influidas por los precios..

```
ods graphics on;
data datos;
      set ejemplos.mercado;
   run;
   proc arima data=datos;
      identify var=ventas(1) crosscorr=precio(1);
      estimate p=1 q=1 input=precio;
      forecast lead=12 interval=month id=date out=resultados;
   run;
```

La parte más importante de la salida es la siguiente:

Estimación por mínimos cuadrados condicional

Parámetro	Estimador	Error estándar	Valor t	Approx Pr > |t|	Retardo	Variable	Desplazamiento
MU	0.48260	0.23012	2.10	0.0386	0	ventas	0
MA1,1	1.00000	0.03003	33.29	<.0001	1	ventas	0
AR1,1	-0.02693	0.10601	-0.25	0.8000	1	ventas	0
NUM1	9.50097	0.23147	41.05	<.0001	0	precio	0

Constante Estimación	0.495596
Varianza Estimación	68.85792
Error std Estimación	8.298067
AIC	703.8393
SBC	714.2197
Número de residuales	99

* AIC y SBC no incluyen determinante de la log.

Correlaciones de los estimadores de parámetro

Parámetro de la variable		ventas MU	ventas MA1,1	ventas AR1,1	precio NUM1
ventas	MU	1.000	-0.096	-0.184	-0.992
ventas	MA1,1	-0.096	1.000	0.174	-0.011
ventas	AR1,1	-0.184	0.174	1.000	0.171
precio	NUM1	-0.992	-0.011	0.171	1.000

Comprobación de los residuales de autocorrelación

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	8.45	4	0.0764	-0.000	0.217	-0.141	-0.073	0.067	-0.067
12	21.15	10	0.0201	-0.027	0.068	0.230	-0.043	0.011	-0.229
18	22.29	16	0.1340	-0.077	0.005	-0.038	-0.044	0.010	0.014
24	28.67	22	0.1544	0.043	0.011	-0.186	-0.065	-0.087	0.036

Comprobación de residuales de correlación cruzada con input precio

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Correlaciones cruzadas					
5	1.91	6	0.9282	-0.010	-0.039	0.094	0.001	-0.068	-0.064
11	9.62	12	0.6491	0.144	0.160	-0.117	-0.082	-0.096	0.044
17	21.87	18	0.2379	0.135	-0.259	0.044	-0.104	0.140	0.078
23	27.90	24	0.2645	0.041	-0.126	0.042	-0.084	0.174	-0.067

Modelo para la variable ventas

Término independiente estimado	0.482598
Periodo(s) de diferenciación	1

Factores autorregresivos

Factor 1: 1 + 0.02693 B**(1)

Factores de la media móvil

Factor 1: 1 - 1 B**(1)

Número de entrada 1

Variable de entrada	precio
Periodo(s) de diferenciación	1
Factor de regresión global	9.50097

Predicciones para la variable ventas

Obs	Predicción	Error std.	95% Límites de confianza	
101	1293.6626	41.4258	1212.4695	1374.8557
102	1304.1870	45.1856	1215.6249	1392.7491
103	1318.1245	48.5607	1222.9472	1413.3017
104	1325.5542	54.6302	1218.4809	1432.6275
105	1335.5260	58.5586	1220.7532	1450.2987
106	1346.1345	62.2089	1224.2072	1468.0617
107	1355.5891	66.0806	1226.0735	1485.1046
108	1365.4841	69.5207	1229.2260	1501.7423
109	1375.4978	72.7914	1232.8293	1518.1663
110	1385.3069	75.9870	1236.3753	1534.2386
111	1395.1923	79.0204	1240.3152	1550.0695
112	1405.0998	81.9401	1244.5001	1565.6994

STATA Y LOS MODELOS ARIMA *(p,d,q)(P,D,Q)s*

Las últimas versiones de STATA incorpora comandos para la identificación, estimación, diagnosis y predicción de modelos ARIMA univariantes de series temporales.

En cuanto a la identificación ya hemos visto que el instrumento esencial son las funciones de autocorrelación y autocorrelación parcial de la serie así como el periodograma y la propia gráfica de la serie. STATA dispone de comandos que realizan estas tareas y que son respectivamente *corrgram, tsline* y *pergram*.

Como ejemplo consideramos la serie mensual del número de pasajeros *air2* de una compañía aérea internacional con la finalidad de identificar un modelo de serie temporal univariante al que se ajuste bien para poder realizar predicciones fiables. En primer lugar graficamos la serie (Figura 7-120).

```
. use http://www.stata-press.com/data/r9/air2
(TIMESLAB: Airline passengers)

. tsline air
```

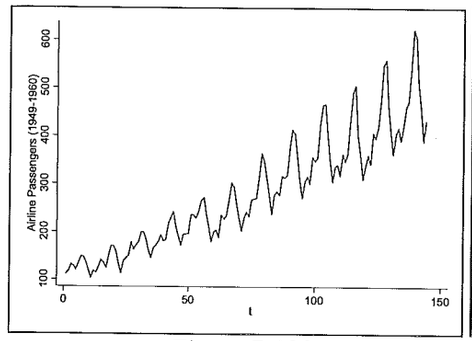


Figura 7-120

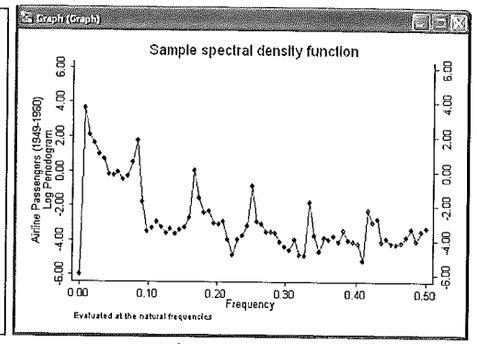


Figura 7-121

Se observa claramente la estacionalidad, que puede corroborarse graficando el periodograma (Figura 7-121) mediante la siguiente sintaxis:

```
. pergram air
```

El periodograma tiene su pico más alto en un punto de abscisa (frecuencia) cercano a 0,08. Luego el periodo será 1/0,08 (aproximadamente 12) y la serie es efectivamente estacional mensual.

A continuación calculamos y graficamos la función de autocorrelación de la serie inicial (Figura 7-122) para analizar su estacionariedad.

```
. corrgram air, lags(20)
```

LAG	AC	PAC	Q	Prob>Q	-1 0 1 [Autocorrelation]	-1 0 1 [Partial Autocor]
1	0.9480	0.9589	132.14	0.0000	|-------	|-------
2	0.8756	-0.3298	245.65	0.0000	|-------	--|
3	0.8067	0.2018	342.67	0.0000	|------	|-
4	0.7526	0.1450	427.74	0.0000	|------	|-
5	0.7138	0.2585	504.8	0.0000	|-----	|--
6	0.6817	-0.0269	575.6	0.0000	|-----	|
7	0.6629	0.2043	643.04	0.0000	|-----	|-

```
8        0.6556   0.1561   709.48   0.0000     |-----          |-
9        0.6709   0.5686   779.59   0.0000     |-----          |----
10       0.7027   0.2926   857.07   0.0000     |-----          |--
11       0.7432   0.8402   944.39   0.0000     |-----          |------
12       0.7604   0.6127   1036.5   0.0000     |------         |----
13       0.7127  -0.6660     1118   0.0000     |-----      -----|
14       0.6463  -0.3846   1185.6   0.0000     |-----       ---|
15       0.5859   0.0787   1241.5   0.0000     |----         |
16       0.5380  -0.0266     1289   0.0000     |----         |
17       0.4997  -0.0581   1330.4   0.0000     |---          |
18       0.4687  -0.0435     1367   0.0000     |---          |
19       0.4499   0.2773   1401.1   0.0000     |---          |--
20       0.4416  -0.0405   1434.1   0.0000     |---          |
```

```
. ac air, lags(20)
```

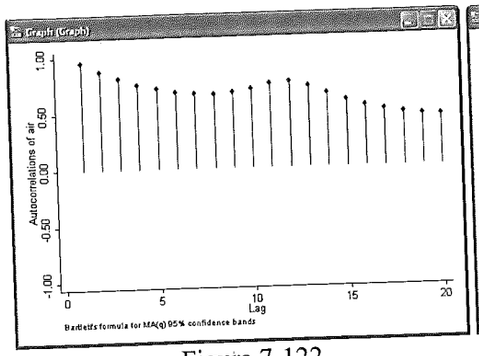


Figura 7-122

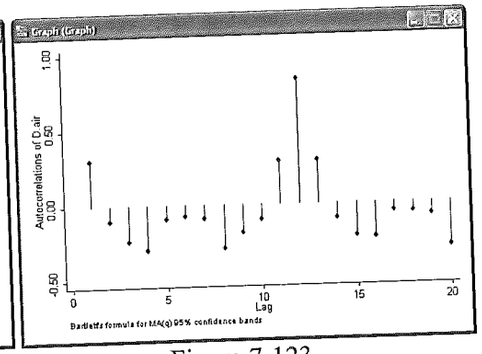


Figura 7-123

Los coeficientes de la función de autocorrelación no decaen rápidamente luego no existe estacionariedad en media y hay que tomar diferencias en la parte regular de la serie original. Ahora graficamos la función de autocorrelación de la serie diferenciada (Figura 7-123) mediante la sintaxis siguiente:

```
. ac D.air,lags(20)
```

Se observa que no hay estacionariedad en la parte estacional porque los términos de la función de autocorrelación de la serie diferenciada no decaen rápidamente y será necesaria también una diferencia estacional. Además, del propio gráfico de la serie se deduce que no hay estacionariedad en varianza y por tanto habrá que tomar logaritmos de la serie inicial. Consideraremos entonces la serie en logaritmos y le aplicamos una diferencia regular y una diferencia estacional de orden 12. Finalmente calcularemos sus funciones de autocorrelación (Figura 7-124) y autocorrelación parcial (Figura 7-125) de la serie en logaritmos doblemente diferenciada.


```
. gen lnair=ln(air)

. ac DS12.D.lnair, lags(20)

. pac DS12.D.lnair, lags(20)
```

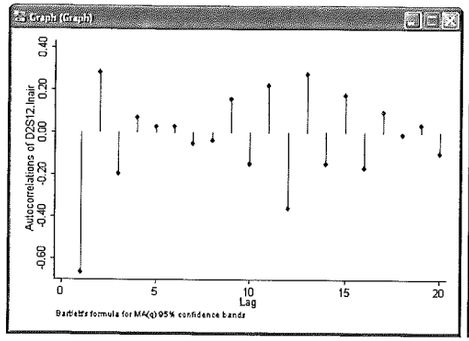


Figura 7-124

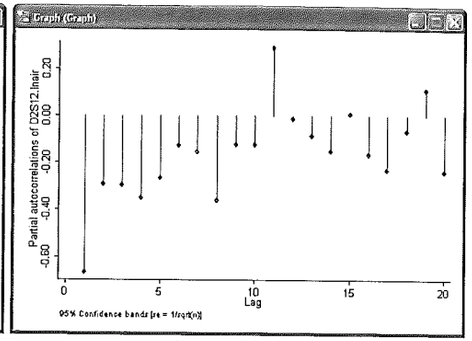


Figura 7-125

Se observa que las funciones de autocorrelación y autocorrelación parcial de la serie en logaritmos con una diferencia regular y otra estacional presentan un patrón MA(1) que se repite para cada periodo estacional. Se concluye entonces que la serie en logaritmos sigue un patrón ARIMA(0,1,1)(0,1,1)$_{12}$.

Una vez identificada la serie, procedemos a su estimación mediante el comando *arima*. Utilizaremos la sintaxis siguiente:

```
. arima lnair, arima(0,1,1) sarima(0,1,1,12) noconstant

(setting optimization to BHHH)
Iteration 0:   log likelihood =   223.8437
Iteration 1:   log likelihood =  239.80405
Iteration 2:   log likelihood =  244.10265
Iteration 3:   log likelihood =  244.65895
Iteration 4:   log likelihood =  244.68945
(switching optimization to BFGS)
Iteration 5:   log likelihood =  244.69431
Iteration 6:   log likelihood =  244.69647
Iteration 7:   log likelihood =  244.69651
Iteration 8:   log likelihood =  244.69651

ARIMA regression

Sample:  14 to 144                              Number of obs      =       131
                                                Wald chi2(2)       =     84.53
Log likelihood =  244.6965                      Prob > chi2        =    0.0000
```

```
------------------------------------------------------------------------------
             |                 OPG
  DS12.lnair |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ARMA         |
          ma |
         L1. |  -.4018324   .0730307    -5.50   0.000    -.5449698   -.2586949
-------------+----------------------------------------------------------------
ARMA12       |
          ma |
         L1. |  -.5569342   .0963129    -5.78   0.000     -.745704   -.3681644
-------------+----------------------------------------------------------------
      /sigma |   .0367167   .0020132    18.24   0.000     .0327708    .0406625
------------------------------------------------------------------------------
```

El modelo estimado resulta ser:

$$(1\text{-B})(1\text{-B})^{12} lnair_t = (10\text{-}0{,}402B)(10\text{-}0{,}557B^{12})a_t.$$

O lo que es lo mismo:

$$\Delta\Delta_{12} lnair_t = -0{,}402\ a_{t-1} - 0{,}557\ a_{t-12} + 0{,}224 a_{t-13} + a_t$$

Suavizado con STATA

STATA permite realizar suavizado exponencial simple de series temporales (*tssmooth exponential*), doble suavizado exponencial (*tssmooth dexponential*), suavizado de Holt-Winters no estacional (*tssmooth hwinters*), suavizado de Holt-Winters estacional (*tssmooth shwinters*) y medias móviles (*tssmooth ma*)

Como ejemplo vamos a realizar predicciones de la serie *air* utilizando el método de alisado de Holt-Winters estacional.

La primera tarea es definir la variable *air* como variable temporal estacional de periodo mensual como sigue:

```
. tsset t, m
        time variable:  t, 1960m2 to 1973m1
```

A continuación ya se puede utilizar el comando de alisado para obtener predicciones para una estación completa (12 meses).

```
. tssmooth shwinters shw1 = air, forecast(12)

computing optimal weights
```



```
Iteration 0:   penalized RSS = -56587.051  (not concave)
Iteration 1:   penalized RSS = -27220.648  (not concave)
Iteration 2:   penalized RSS = -26929.075  (not concave)
Iteration 3:   penalized RSS = -26927.738  (not concave)
Iteration 4:   penalized RSS = -26926.926  (not concave)
Iteration 5:   penalized RSS = -17496.526  (not concave)
Iteration 6:   penalized RSS = -16692.431  (not concave)
Iteration 7:   penalized RSS = -16565.013
Iteration 8:   penalized RSS =  -16531.28  (backed up)
Iteration 9:   penalized RSS =  -16514.52
Iteration 10:  penalized RSS = -16514.493
Iteration 11:  penalized RSS = -16514.493

Optimal weights:
                              alpha = 0.2891
                               beta = 0.0212
                              gamma = 0.8579
penalized sum-of-squared residuals = 16514.49
          sum-of-squared residuals = 16514.49
          root mean squared error  = 10.70906
```

Si queremos representar la serie inicial junto con la serie alisada y las predicciones en el mismo gráfico, utilizaremos la sintaxis siguiente:

```
. line air shw1 t, title("Predición estacional de Holt-Winters")
```

Se obtiene el gráfico de la Figura 7-126.

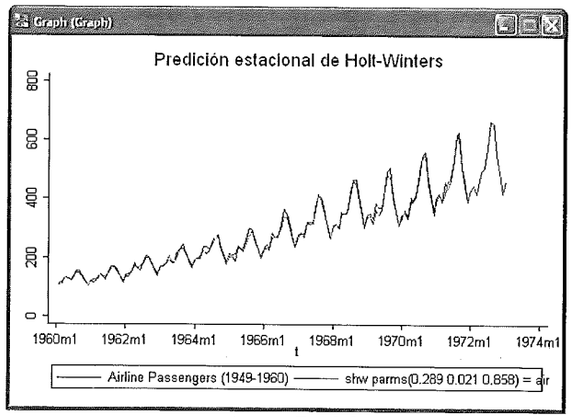


Figura 7-126

Para ver la serie alisada y las predicciones, se lista la variable *shw1* como sigue:

```
. list shw1

     +----------+
     |     shw1 |
     |----------|
  1. | 107.0562 |
  2. | 111.8558 |
  3. | 131.8469 |
  4. | 126.9532 |
  5. | 126.5536 |
     |----------|
  6. | 140.2717 |
  7. | 154.1386 |
  8. | 154.1928 |
  9. | 137.3746 |
 10. | 121.0375 |
     |----------|
 11. | 106.2217 |
 12. | 121.2639 |
 13. | 126.5788 |
 14. |  127.308 |
 15. | 142.2327 |
     |----------|
 16. | 137.3846 |
 17. | 130.2811 |
.................................
```

Ejercicio 7-1. Como primer ejemplo ajustamos un modelo de la función de transferencia con intervenciones para el período posterior a 1960 en los meses de verano y en los meses de invierno, para una serie de datos mensuales relativos a la cantidad de ozono en el aire.

Se trata de ajustar un modelo ARIMA a los datos dados para la concentración de ozono en el aire teniendo presente que en esta variable interviene la estación del año, en concreto las estaciones veraniegas y las estaciones invernales. Utilizaremos entonces un modelo con intervención en verano y en invierno. La sintaxis SAS adecuada podría ser la siguiente:

```
title1 'Analisis de la intervencion para la cocentracion de ozono';
title2 '(Box and Tiao, JASA 1975 P.70)';
data aire;
   input ozono @@;
   label ozono  = 'Concentracion de ozono'
         x1     = 'Intervencion para el periodo posterior a 1960'
         summer = 'Intervencion en los meses de verano'
         winter = 'Intervencion en los meses de invierno';
   date = intnx( 'month', '31dec1954'd, _n_ );
   format date monyy.;
   month = month( date );
   year = year( date );
   x1 = year >= 1960;
   summer = ( 5 < month < 11 ) * ( year > 1965 );
   winter = ( year > 1965 ) - summer;
datalines;
```



```
   2.7  2.0  3.6  5.0  6.5  6.1  5.9  5.0  6.4  7.4  8.2  3.9
   4.1  4.5  5.5  3.8  4.8  5.6  6.3  5.9  8.7  5.3  5.7  5.7
   3.0  3.4  4.9  4.5  4.0  5.7  6.3  7.1  8.0  5.2  5.0  4.7
   3.7  3.1  2.5  4.0  4.1  4.6  4.4  4.2  5.1  4.6  4.4  4.0
   2.9  2.4  4.7  5.1  4.0  7.5  7.7  6.3  5.3  5.7  4.8  2.7
   1.7  2.0  3.4  4.0  4.3  5.0  5.5  5.0  5.4  3.8  2.4  2.0
   2.2  2.5  2.6  3.3  2.9  4.3  4.2  4.2  3.9  3.9  2.5  2.2
   2.4  1.9  2.1  4.5  3.3  3.4  4.1  5.7  4.8  5.0  2.8  2.9
   1.7  3.2  2.7  3.0  3.4  3.8  5.0  4.8  4.9  3.5  2.5  2.4
   1.6  2.3  2.5  3.1  3.5  4.5  5.7  5.0  4.6  4.8  2.1  1.4
   2.1  2.9  2.7  4.2  3.9  4.1  4.6  5.8  4.4  6.1  3.5  1.9
   1.8  1.9  3.7  4.4  3.8  5.6  5.7  5.1  5.6  4.8  2.5  1.5
   1.8  2.5  2.6  1.8  3.7  3.7  4.9  5.1  3.7  5.4  3.0  1.8
   2.1  2.6  2.8  3.2  3.5  3.5  4.9  4.2  4.7  3.7  3.2  1.8
   2.0  1.7  2.8  3.2  4.4  3.4  3.9  5.5  3.8  3.2  2.3  2.2
   1.3  2.3  2.7  3.3  3.7  3.0  3.8  4.7  4.6  2.9  1.7  1.3
   1.8  2.0  2.2  3.0  2.4  3.5  3.5  3.3  2.7  2.5  1.6  1.2
   1.5  2.0  3.1  3.0  3.5  3.4  4.0  3.8  3.1  2.1  1.6  1.3
   ;
   proc arima data=aire;

      /*--- Identificacion con una diferencia estacional e
        intervención para la serie del ozono ---*/
      identify var=ozono(12) crosscorr=( x1(12) summer winter )
noprint;

      /*--- Ajuste a un modelo ARIMA(0,0,1)(0,1,1)12 ---*/
      /*---  por el metodo de maxima verosimilitiud ---*/
      estimate q=(1)(12) input=( x1 summer winter )
               noconstant method=ml itprint;

      /*--- 12 Predicciones ---*/
      forecast  lead=12 id=date interval=month;

   run;
```

La salida es la siguiente:

```
Analisis de la intervencion para la cocentracion de ozono
(Box and Tiao, JASA 1975 P.70)
The ARIMA Procedure

Preliminary Estimation

Initial Moving Average
Estimates

                Estimate
 1               -0.29241

Initial Moving Average
Estimates

                Estimate
 12               0.40740

White Noise Variance Est    0.944969
Conditional Least Squares Estimation

Iteration      SSE     MA1,1     MA2,1     NUM1      NUM2      NUM3    Lambda    R Crit
        0   154.53  -0.29241   0.40740  -1.13490  -0.11731   0.05581   0.00001         1
        1   146.20  -0.29256   0.59844  -1.20292  -0.29784  -0.11572      1E-6  0.230552
        2   145.88  -0.30071   0.59239  -1.26173  -0.26252  -0.08247      1E-7  0.046601
        3   145.88  -0.29976   0.59242  -1.26246  -0.26150  -0.08197      1E-8  0.001345
        4   145.88  -0.29983   0.59234  -1.26243  -0.26154  -0.08196      1E-9  0.000125
```

Maximum Likelihood Estimation

Iter	Loglike	MA1,1	MA2,1	NUM1	NUM2	NUM3	Lambda	R Crit
0	-249.07778	-0.29983	0.59234	-1.26243	-0.26154	-0.08196	0.00001	1
1	-245.89135	-0.26830	0.76634	-1.34490	-0.23984	-0.07578	1E-6	0.169445
2	-245.88484	-0.26653	0.76623	-1.33046	-0.23939	-0.08025	1E-7	0.008044
3	-245.88482	-0.26689	0.76661	-1.33070	-0.23936	-0.08020	1E-8	0.000603
4	-245.88481	-0.26684	0.76665	-1.33062	-0.23936	-0.08021	1E-9	0.000073

ARIMA Estimation Optimization Summary

Estimation Method	Maximum Likelihood
Parameters Estimated	5
Termination Criteria	Maximum Relative Change in Estimates
Iteration Stopping Value	0.001
Criteria Value	0.000195
Alternate Criteria	Relative Change in Objective Function
Alternate Criteria Value	1.247E-8
Maximum Absolute Value of Gradient	0.00712
R-Square Change from Last Iteration	0.000073
Objective Function	Log Gaussian Likelihood
Objective Function Value	-245.885
Marquardt's Lambda Coefficient	1E-9
Numerical Derivative Perturbation Delta	0.001
Iterations	4

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	-0.26684	0.06710	-3.98	<.0001	1	ozono	0
MA2,1	0.76665	0.05973	12.83	<.0001	12	ozono	0
NUM1	-1.33062	0.19236	-6.92	<.0001	0	x1	0
NUM2	-0.23936	0.05952	-4.02	<.0001	0	summer	0
NUM3	-0.08021	0.04978	-1.61	0.1071	0	winter	0

Variance Estimate	0.634506
Std Error Estimate	0.796559
AIC	501.7696
SBC	518.3602
Number of Residuals	204

Correlations of Parameter Estimates

Variable Parameter		ozono MA1,1	ozono MA2,1	x1 NUM1	summer NUM2	winter NUM3
ozono	MA1,1	1.000	0.090	-0.039	0.062	-0.034
ozono	MA2,1	0.090	1.000	-0.169	0.211	0.022
x1	NUM1	-0.039	-0.169	1.000	-0.124	-0.107
summer	NUM2	0.062	0.211	-0.124	1.000	0.097
winter	NUM3	-0.034	0.022	-0.107	0.097	1.000

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	7.47	4	0.1132	0.017	0.054	0.043	0.101	-0.022	0.140
12	10.21	10	0.4220	-0.024	-0.059	-0.047	0.014	0.032	0.072
18	14.53	16	0.5593	0.054	0.006	-0.110	0.028	-0.042	0.043
24	19.99	22	0.5834	0.003	-0.074	-0.074	0.098	-0.038	0.043
30	27.00	28	0.5180	-0.072	-0.035	0.023	-0.028	-0.107	0.100
36	32.65	34	0.5336	0.022	-0.099	-0.006	0.087	-0.046	0.053

Model for variable ozono

Period(s) of Differencing 12

No mean term in this model.
Moving Average Factors

Factor 1: 1 + 0.26684 B**(1)
Factor 2: 1 - 0.76665 B**(12)

Input Number 1

Input Variable	x1
Period(s) of Differencing	12
Overall Regression Factor	-1.33062

Input Number 2

Input Variable	summer
Overall Regression Factor	-0.23936

Input Number 3

Input Variable	winter
Overall Regression Factor	-0.08021

El ajuste del modelo es muy Bueno porque los p-valores de todos los parámetros estimados son muy bajos. Vistas las estimaciones de los parámetros, el modelo de serie temporal ajustado resulta ser el siguiente:

$$(1\text{-}B^{12})ozono_t = (1 + 0.28884\ B)(1 - 0.78885\ B^{12})\ a_t + Intervención$$

Ejercicio 7-2: El fichero empresa.txt contiene los datos sobre fechas precios y ventas de una empresa. Se pide:

a) Ajustar razonadamente la serie de ventas a un modelo ARIMA del tipo:

$$(1-B)Y_t = \mu + \frac{(1-\theta_1 B)}{(1-\phi_1 B)} a_t$$

b) Suponiendo que la serie de ventas es mensual, realizar un ajuste a un modelo ARIMA estacional del tipo:

$$(1-B)(1-B^{12})Y_t = \mu + \frac{(1-\theta_{1,1}B-\theta_{1,2}B^2)}{(1-\phi_{1,1}B)}\frac{(1-\theta_{2,1}B^{12})}{(1-\phi_{2,1}B^{12})} a_t$$

c) Considerando las series de ventas Yt y de precios Xt, realizar un ajuste a un modelo de la función de transferencia del tipo:

$$Y_t = \mu + \frac{\varpi_0 - \varpi_1 B - \varpi_2 B^2 - \varpi_3 B^3}{(1-\delta_1 B)} X_t + a_t$$

En el primer apartado se está pidiendo un ajuste a un modelo del tipo ARIMA(1,1,1). Para ejecutar las fases de identificación, estimación y predicción se utilizará la siguiente sintaxis SAS:

```
data empresa;
infile 'c:\sas\empresa.txt';
input tiempo precio ventas;
proc arima data=empresa;
      identify var=ventas(1) nlag=10;
      run;
      estimate p=1 q=1;
      run;
      forecast lead=12 id=tiempo;
run;
```

La salida es la siguiente:

```
                    The ARIMA Procedure

                 Name of Variable = ventas

Period(s) of Differencing                                  1
Mean of Working Series                              9.984535
Standard Deviation                                  49.09608
Number of Observations                                    99
Observation(s) eliminated by differencing                  1

                      Autocorrelations

Lag   Covariance   Correlation   -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1     Std Error
  0     2410.425       1.00000   |                    |********************|            0
  1    -1126.910       -.46752   |            *********|   .                |     0.100504
  2   -67.527104       -.02801   |                .   *|    .               |     0.120485
  3   -59.259537       -.02458   |                .    |    .               |     0.120551
  4    72.821057       0.03021   |                .    |*   .               |     0.120601
  5      169.375       0.07027   |                .    |*   .               |     0.120678
  6     -246.894       -.10243   |                .  **|    .               |     0.121090
  7     -224.587       -.09317   |                .  **|    .               |     0.121962
  8      436.829       0.18123   |                .    |****.               |     0.122679
  9     -149.552       -.06204   |                .   *|    .               |     0.125354
 10   -55.289321       -.02294   |                .    |    .               |     0.125664

                  "." marks two standard errors
                    Inverse Autocorrelations

Lag   Correlation   -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
  1      0.84798    |                .   |*****************   |
  2      0.72043    |                .   |**************      |
  3      0.61291    |                .   |************        |
  4      0.50129    |                .   |**********          |
  5      0.41206    |                .   |********            |
  6      0.34884    |                .   |*******             |
  7      0.25450    |                .   |*****               |
  8      0.14229    |                .   |***.                |
  9      0.08420    |                .   |** .                |
 10      0.03482    |                .   |*  .                |

                    Partial Autocorrelations

Lag   Correlation   -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
  1     -0.46752    |            *********|   .                |
  2     -0.31556    |               ******|   .                |
  3     -0.26910    |                *****|   .                |
  4     -0.19349    |                 ****|   .                |
  5     -0.03648    |                  . *|   .                |
  6     -0.10361    |                  . **|   .               |
  7     -0.27830    |               ******|   .                |
  8     -0.09341    |                  . **|   .               |
  9     -0.09932    |                  . **|   .               |
 10     -0.13218    |                  .***|   .               |

                Autocorrelation Check for White Noise

To     Chi-          Pr >
Lag    Square   DF   ChiSq    ----------------Autocorrelations----------------

 6     24.19     6   0.0005   -0.468   -0.028   -0.025    0.030    0.070   -0.102

                ARIMA Estimation Optimization Summary

Estimation Method                                  Conditional Least Squares
Parameters Estimated                                                       3
Termination Criteria                     Maximum Relative Change in Estimates
Iteration Stopping Value                                               0.001
Criteria Value                                                       0.25687
Maximum Absolute Value of Gradient                                  54129.14
```



```
R-Square Change from Last Iteration                                 0.149317
Objective Function                                  Sum of Squared Residuals
Objective Function Value                                            125835.9
Marquardt's Lambda Coefficient                                       0.00001
Numerical Derivative Perturbation Delta                                0.001
Iterations                                                                26
Warning Message                          Estimates may not have converged.

                  Conditional Least Squares Estimation

                                Standard                    Approx
Parameter      Estimate           Error    t Value     Pr > |t|     Lag

MU              9.87916         0.07998     123.52       <.0001       0
MA1,1           1.00000         0.03253      30.74       <.0001       1
AR1,1           0.10339         0.10396       0.99       0.3224       1

                Constant Estimate       8.857725
                Variance Estimate       1310.791
                Std Error Estimate      36.20485
                AIC                     994.5636
                SBC                     1002.349
                Number of Residuals           99
            * AIC and SBC do not include log determinant.

                  Correlations of Parameter Estimates

Parameter        MU     MA1,1     AR1,1

MU            1.000    -0.475    -0.063
MA1,1        -0.475     1.000     0.185
AR1,1        -0.063     0.185     1.000

                  Autocorrelation Check of Residuals

To     Chi-          Pr >
Lag    Square   DF   ChiSq    ----------------Autocorrelations----------------

 6      2.44     4   0.6555   -0.051   -0.012   -0.017    0.030    0.025   -0.136
12      5.53    10   0.8529   -0.099    0.130   -0.019   -0.020    0.019    0.019
18      9.60    16   0.8865   -0.150    0.003    0.022   -0.048    0.091    0.027
24     14.19    22   0.8945    0.048    0.054    0.040    0.135   -0.099   -0.026

                      Model for variable ventas

                Estimated Mean                9.879157
                Period(s) of Differencing            1

                     Autoregressive Factors

                Factor 1:  1 - 0.10339 B**(1)

                     Moving Average Factors

                Factor 1:  1 - 1 B**(1)

                  Forecasts for variable ventas

Obs      Forecast    Std Error      95% Confidence Limits

101     1298.0753      36.2048    1227.1151    1369.0354
102     1306.9873      36.3978    1235.6488    1378.3258
103     1316.7664      36.3999    1245.4239    1388.1090
104     1326.6353      36.3999    1255.2927    1397.9778
105     1336.5134      36.3999    1265.1708    1407.8559
106     1346.3924      36.3999    1275.0498    1417.7349
107     1356.2715      36.3999    1284.9290    1427.6141
108     1366.1507      36.3999    1294.8082    1437.4933
109     1376.0299      36.3999    1304.6873    1447.3724
110     1385.9090      36.3999    1314.5665    1457.2516
111     1395.7882      36.3999    1324.4456    1467.1307
112     1405.6673      36.3999    1334.3248    1477.0099
```

En cuanto a la fase de identificación, se han obtenido funciones de autocorrelación adecuadas a un modelo ARIMA(1,1,1). En cuanto a la fase de estimación (p-valores pequeños), se ha obtenido la siguiente ecuación de ajuste para el modelo:

$$(1-B)Y_t = 9{,}879157 + \frac{(1-B)}{(1-0{,}102039B)}a_t$$

En cuanto a la fase de diagnosis, el chequeo de los residuos indica que no hay problemas par aceptar el modelo. Las predicciones se ven al final de la salida.

En la segunda parte del problema se trata de modelizar la serie de ventas mediante un modelo estacional ARIMA(1,1,2)(1,1,1)$_{12}$. Utilizaremos la siguiente sintaxis SAS:

```
data empresa;
infile 'c:\sas\empresa.txt';
input tiempo precio ventas;
proc arima data=empresa;
      identify var=ventas(1,12) nlag=10;
      run;
      estimate p=(1)(12) q=(1,2)(12);
      run;
      forecast lead=12 id=tiempo interval=month ;
run;
```

La salida es la siguiente:

The ARIMA Procedure

Name of Variable = ventas

Period(s) of Differencing	1,12
Mean of Working Series	-0.18971
Standard Deviation	66.94818
Number of Observations	87
Observation(s) eliminated by differencing	13

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	4482.059	1.00000	| |********** |	0
1	-1901.163	-.42417	| ********| . |	0.107211
2	-231.464	-.05164	| . *| . |	0.125022
3	-214.301	-.04781	| . *| . |	0.125266
4	99.079200	0.02211	| . | . |	0.125476
5	338.331	0.07549	| . |** . |	0.125521
6	-154.866	-.03455	| . *| . |	0.126042
7	-629.084	-.14036	| . ***| . |	0.126150
8	606.965	0.13542	| . |*** . |	0.127933
9	39.707086	0.00886	| . | . |	0.129570
10	-589.119	-.13144	| . ***| . |	0.129577

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.78648	| . |********** |
2	0.65803	| . |********** |
3	0.56671	| . |********** |
4	0.44980	| . |********* |
5	0.36176	| . |******* |
6	0.32627	| . |******* |
7	0.27219	| . |***** |
8	0.16564	| . |***. |
9	0.11584	| . |** . |
10	0.07507	| . |** . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.42417	| ********| . |
2	-0.28237	| ******| . |
3	-0.25928	| *****| . |
4	-0.19781	| ****| . |
5	-0.05556	| . *| . |
6	-0.03985	| . *| . |
7	-0.21881	| ****| . |
8	-0.07603	| . **| . |
9	-0.02560	| . *| . |
10	-0.21383	| ****| . |

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	17.35	6	0.0081	-0.424	-0.052	-0.048	0.022	0.075	-0.035

Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	-0.53501	0.32466	-1.65	0.1032	0
MA1,1	0.71315	1.99717	0.36	0.7220	1
MA1,2	0.14837	1.73384	0.09	0.9320	2
MA2,1	0.63187	0.16976	3.72	0.0004	12
AR1,1	-0.10828	2.00963	-0.05	0.9572	1
AR2,1	-0.03082	0.20320	-0.15	0.8798	12

Constant Estimate	-0.61121
Variance Estimate	2172.13
Std Error Estimate	46.60611
AIC	921.1397
SBC	935.9351
Number of Residuals	87

* AIC and SBC do not include log determinant.

Correlations of Parameter Estimates

Parameter	MU	MA1,1	MA1,2	MA2,1	AR1,1	AR2,1
MU	1.000	-0.070	0.064	-0.282	-0.068	-0.193
MA1,1	-0.070	1.000	-0.999	0.034	0.998	0.076
MA1,2	0.064	-0.999	1.000	-0.036	-0.998	-0.076
MA2,1	-0.282	0.034	-0.036	1.000	0.031	0.812
AR1,1	-0.068	0.998	-0.998	0.031	1.000	0.071
AR2,1	-0.193	0.076	-0.076	0.812	0.071	1.000

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	1.76	1	0.1840	-0.002	-0.024	-0.067	0.098	0.034	-0.055
12	5.51	7	0.5985	-0.050	0.166	-0.050	-0.059	0.042	0.001
18	9.90	13	0.7018	-0.194	0.012	0.018	0.009	0.036	0.048
24	14.47	19	0.7554	0.094	0.094	-0.031	0.119	-0.076	0.017

Model for variable ventas

Estimated Mean -0.53501
Period(s) of Differencing 1,12

Autoregressive Factors

Factor 1: 1 + 0.10828 B**(1)
Factor 2: 1 + 0.03082 B**(12)

Moving Average Factors

Factor 1: 1 - 0.71315 B**(1) - 0.14837 B**(2)
Factor 2: 1 - 0.63187 B**(12)

Forecasts for variable ventas

Obs	Forecast	Std Error	95% Confidence Limits	
101	1260.5616	46.6061	1169.2153	1351.9079
102	1287.0868	47.3433	1194.2956	1379.8780
103	1308.6390	47.6678	1215.2117	1402.0662
104	1294.2210	48.0258	1200.0922	1388.3498
105	1310.9603	48.3772	1216.1428	1405.7778
106	1302.2069	48.7264	1206.7049	1397.7089
107	1377.6882	49.0731	1281.5066	1473.8698
108	1339.0944	49.4174	1242.2380	1435.9508
109	1338.4927	49.7594	1240.9662	1436.0193
110	1353.9954	50.0989	1255.8033	1452.1875
111	1353.3511	50.4362	1254.4979	1452.2043
112	1365.9624	50.7713	1266.4525	1465.4723

El ajuste al modelo ARIMA(1,1,2)(1,1,1)$_{12}$ supera las fases de identificación y diagnosis, resultando como modelo estimado el siguiente:

$$\boxed{(1-B)(1-B^{12})Y_t = -0{,}53501 + \frac{(1-0{,}71315B-0{,}14837B^2)}{(1+0{,}10828B)}\frac{(1-0{,}63187B^{12})}{(1+0{,}03082B^{12})}a_t}$$

Las predicciones según este modelo se observan al final de la salida.

Para realizar el ajuste al modelo de la función de transferencia del tercer apartado del problema de utilizará la siguiente sintaxis SAS:

```
proc arima data=empresa;
     identify var=ventas crosscorr=precio;
     estimate input=( (1 2 3) / (1) precio );
     run;
```

La salida es la siguiente:

The ARIMA Procedure

Name of Variable = ventas

Mean of Working Series 796.6085
Standard Deviation 288.2731
Number of Observations 100

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	495.11	6	<.0001	0.956	0.926	0.898	0.870	0.843	0.804
12	829.75	12	<.0001	0.774	0.748	0.717	0.687	0.659	0.630
18	1035.66	18	<.0001	0.603	0.575	0.547	0.518	0.491	0.457
24	1142.62	24	<.0001	0.435	0.408	0.383	0.354	0.327	0.305

Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	-199.62193	3.27412	-60.97	<.0001	0	ventas	0
NUM1	9.62530	0.21488	44.79	<.0001	0	precio	0
NUM1,1	0.60670	24.49979	0.02	0.9803	1	precio	0
NUM1,2	-0.27985	0.22812	-1.23	0.2231	2	precio	0
NUM1,3	-0.06470	0.82479	-0.08	0.9376	3	precio	0
DEN1,1	0.06360	2.54135	0.03	0.9801	1	precio	0

Constant Estimate -199.622
Variance Estimate 70.88881
Std Error Estimate 8.419549
AIC 694.4084
SBC 709.8567
Number of Residuals 97
* AIC and SBC do not include log determinant.

Correlations of Parameter Estimates

Variable Parameter		ventas MU	precio NUM1	precio NUM1,1	precio NUM1,2	precio NUM1,3	precio DEN1,1
ventas	MU	1.000	-0.203	0.174	0.057	0.146	0.173
precio	NUM1	-0.203	1.000	-0.327	0.064	-0.249	-0.331
precio	NUM1,1	0.174	-0.327	1.000	0.299	0.969	1.000
precio	NUM1,2	0.057	0.064	0.299	1.000	0.186	0.303
precio	NUM1,3	0.146	-0.249	0.969	0.186	1.000	0.969
precio	DEN1,1	0.173	-0.331	1.000	0.303	0.969	1.000

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	8.43	6	0.2084	-0.029	0.230	-0.138	-0.081	0.053	-0.029
12	21.09	12	0.0490	-0.027	0.089	0.203	-0.072	0.025	-0.241
18	22.28	18	0.2198	-0.069	-0.017	-0.037	-0.058	-0.004	-0.019
24	30.09	24	0.1817	0.005	-0.028	-0.212	-0.072	-0.094	0.042

Model for variable ventas

Estimated Intercept -199.622

Input Number 1

Input Variable precio

Numerator Factors

Factor 1: 9.6253 - 0.6067 B**(1) + 0.27985 B**(2) + 0.0647 B**(3)

Denominator Factors

Factor 1: 1 - 0.0636 B**(1)

El modelo de función de transferencia identificado, estimado y diagnosticado correctamente es el siguiente:

$$\boxed{Y_t = -199{,}622 + \frac{9{,}6253 - 0{,}6067B + 0{,}27985B^2 + 0{,}0647B^3}{(1 - 0{,}0636B)} X_t + a_t}$$

Ejercicio 7-3. Ajustar un modelo de función de transferencia que explique el nivel de CO_2 en la atmósfera (variable Y) en función de la tasa de uso de gas para calderas de calefacción (variable X). Los datos de ambas series fueron recogidos por Box y Jenkins en 1976 (Series J) y se observan en la sintaxis de entrada del procedimiento SAS que resuelve el problema.

Para identificar y estimar el modelo de función de transferencia adecuado comenzamos realizando una modelización ARIMA univariante de la variable independiente X. A continuación se utilizan las correlaciones cruzadas de la variable X y la variable Y. Después de ajustar un modelo para X, se preblanquean X e Y por este modelo antes de computar las correlaciones cruzadas. A continuación se ajusta un modelo de la función de transferencia sin estructura de ruido blanco y los residuos de este modelo se identifican mediante la opción PLOT. Por último, se ajusta el modelo completo de la función de transferencia con término de ruido.

La sintaxis SAS adecuada, a través del procedimiento ARIMA sería la siguiente:

```
data datos;
set ejemplos.seriesj;
run;

proc arima data=datos;
   /*--- Se identifica la variable independiente X---*/
   identify var=x nlag=10;
   run;
   /*--- Se ajusta un modelo ARIMA univariante para X ---*/
   estimate p=3;
   run;
   /*-Correlaciones cruzadas y preblanqueado de las series-*/
   identify var=y crosscorr=(x) nlag=10;
   run;
   /*--- Ajuste de la función de transferencia - observar los
   residuos ---*/
   estimate input=( 3$ (1,2)/(1,2) x ) plot;
   run;
   /*--- Estimación del modelo completo --*/
   estimate p=2 input=( 3$ (1,2)/(1) x );
   run;
```


La salida comienza presentando la ***identificación de X como un AR(3)***

```
                                   Proc ARIMA

                              Nombre de la variable = x

                        Media de series de trabajo     -0.05683
                        Desviación estándar            1.070952
                        Número de observaciones             296

                                  Autocorrelaciones

Retardo   Covarianza   Correlación   -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1   Error Std.

      0     1.146938       1.00000   |                    |********************|          0
      1     1.092430       0.95247   |                  . |******************** |   0.058124
      2     0.956652       0.83409   |                .   |*****************    |   0.097510
      3     0.782051       0.68186   |              .     |**************       |   0.119201
      4     0.609291       0.53123   |              .     |***********          |   0.131721
      5     0.467380       0.40750   |             .      |********             |   0.138770
      6     0.364957       0.31820   |             .      |******               |   0.142756
      7     0.298427       0.26019   |             .      |*****.               |   0.145132
      8     0.260943       0.22751   |             .      |*****.               |   0.146699
      9     0.244378       0.21307   |             .      |**** .               |   0.147887
     10     0.238942       0.20833   |             .      |**** .               |   0.148920

                        "." marca dos errores estándar

                           Autocorrelaciones inversas

                Retardo   Correlación   -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1

                      1      -0.71090   |      **************| .                 |
                      2       0.26217   |                 . |*****              |
                      3      -0.13005   |                ***| .                 |
                      4       0.14777   |                 . |***                |
                      5      -0.06803   |                 .*| .                 |
                      6      -0.01147   |                 . | .                 |
                      7      -0.01649   |                 . | .                 |
                      8       0.06108   |                 . |*.                 |
                      9      -0.04490   |                 .*| .                 |
                     10       0.01100   |                 . | .                 |

                           Autocorrelaciones parciales

                Retardo   Correlación   -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1

                      1       0.95247   |                 . |******************* |
                      2      -0.78796   |    ***************| .                 |
                      3       0.33897   |                 . |*******            |
                      4       0.12121   |                 . |**                 |
                      5       0.05896   |                 . |*.                 |
                      6      -0.11147   |                **| .                 |
                      7       0.04862   |                 . |*.                 |
                      8       0.09945   |                 . |**                 |
                      9       0.01587   |                 . | .                 |
                     10      -0.06973   |                 .*| .                 |

                        Comprobación de autocorrelación del ruido blanco

   Para      Chi-          Pr >
retardo   cuadrado   DF    ChiSq   ------------------Autocorrelaciones------------------

      6     786.35    6   <.0001    0.952    0.834    0.682    0.531    0.408    0.318
```

A continuación, la salida muestra la ***estimación de un modelo AR(3) para X***

Estimación por mínimos cuadrados condicional

Parámetro	Estimación	Error estándar	Valor t	Aprox Pr > |t|	Retardo
MU	-0.12280	0.10902	-1.13	0.2609	0
AR1,1	1.97607	0.05499	35.94	<.0001	1
AR1,2	-1.37499	0.09967	-13.80	<.0001	2
AR1,3	0.34336	0.05502	6.24	<.0001	3

Constante Estimación	-0.00682
Varianza Estimación	0.035797
Error estand Estimación	0.1892
AIC	-141.667
SBC	-126.906
Número de residuales	296

* AIC y SBC no incluyen determinante de la log.

Correlations of Parameter Estimates

Parámetro	MU	AR1,1	AR1,2	AR1,3
MU	1.000	-0.017	0.014	-0.016
AR1,1	-0.017	1.000	-0.941	0.790
AR1,2	0.014	-0.941	1.000	-0.941
AR1,3	-0.016	0.790	-0.941	1.000

Autocorrelation Check of Residuals

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	10.30	3	0.0162	-0.042	0.068	0.056	-0.145	-0.009	0.059
12	19.89	9	0.0186	0.014	0.002	-0.055	0.035	0.143	-0.079
18	27.92	15	0.0221	0.099	0.043	-0.082	0.017	0.066	-0.052
24	31.05	21	0.0729	-0.078	0.024	0.015	0.030	0.045	0.004
30	34.58	27	0.1499	-0.007	-0.004	0.073	-0.038	-0.062	0.003
36	38.84	33	0.2231	0.010	0.002	0.082	0.045	0.056	-0.023
42	41.18	39	0.3753	0.002	0.033	-0.061	-0.003	-0.006	-0.043
48	42.73	45	0.5687	0.018	0.051	-0.012	0.015	-0.027	0.020

Modelo para la variable x

Media estimada -0.1228

Factores autoregresivos

Factor 1: 1 - 1.97607 B**(1) + 1.37499 B**(2) - 0.34336 B**(3)

La serie X se ha identificado y estimado adecuadamente como un AR(3). La diagnosis del modelo es correcta (p-valores pequeños de los parámetros estimados, salvo quizá la media) y contraste de aleatoriedad residual correcto.

A continuación, ***se preblanquean Y y X por el modelo ajustado y se calculan las autocorrelaciones cruzadas.***

Nombre de la variable = y

Media de series de trabajo	53.50912
Desviación estándar	3.196707
Número de observaciones	296

Autocorrelaciones

Retardo	Covarianza	Correlación	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Error Std.
0	10.218937	1.00000	| |********** |	0
1	9.920101	0.97076	| . |********** |	0.058124
2	9.156572	0.89604	| . |********** |	0.098720
3	8.099002	0.79255	| . |********** |	0.123169
4	6.948508	0.67996	| . |********** |	0.139337
5	5.870550	0.57448	| . |********** |	0.150129
6	4.960762	0.48545	| . |********** |	0.157381
7	4.251890	0.41608	| . |******** |	0.162361
8	3.736119	0.36561	| . |******* |	0.165924
9	3.376155	0.33038	| . |******* |	0.168623
10	3.132316	0.30652	| . |******. |	0.170796

"." marca dos errores estándar

Autocorrelaciones inversas

Retardo	Correlación	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.60406	| **********| . |
2	0.03304	| . |*. |
3	0.05747	| . |*. |
4	0.04795	| . |*. |
5	-0.01833	| . | . |
6	-0.04534	| .*| . |
7	0.05122	| . |*. |
8	-0.03710	| .*| . |
9	0.02130	| . | . |
10	-0.00607	| . | . |

Autocorrelaciones parciales

Retardo	Correlación	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.97076	| . |********** |
2	-0.80388	| **********| . |
3	0.18833	| . |**** |
4	0.25999	| . |***** |
5	0.05949	| . |*. |
6	-0.06258	| .*| . |
7	-0.01435	| . | . |
8	0.05490	| . |*. |
9	0.00545	| . | . |
10	0.03141	| . |*. |

Comprobación de autocorrelación del ruido blanco

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	1023.15	6	<.0001	0.971	0.896	0.793	0.680	0.574	0.485

Correlación de y y x

Número de observaciones	296
Varianza de las series transformadas y	0.131438
Varianza de las series transformadas x	0.035357

Se ha suprimido el ruido blanco de ambas series.

Correlaciones cruzadas

Retardo	Covarianza	Correlación	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
-10	0.0015683	0.02301	| . | . |
-9	0.00013502	0.00198	| . | . |
-8	-0.0060480	-.08872	| **| . |

-7	-0.0017624	-.02585	.*| .
-6	-0.0080539	-.11814	**| .
-5	-0.0000944	-.00138	. | .
-4	-0.0012802	-.01878	. | .
-3	-0.0031078	-.04559	.*| .
-2	0.00065212	0.00957	. | .
-1	-0.0019166	-.02811	.*| .
0	-0.0003673	-.00539	. | .
1	0.0038939	0.05712	. |*.
2	-0.0016971	-.02489	. | .
3	-0.019231	-.28210	******| .
4	-0.022479	-.32974	*******| .
5	-0.030909	-.45341	*********| .
6	-0.018122	-.26583	*****| .
7	-0.011426	-.16761	***| .
8	-0.0017355	-.02546	.*| .
9	0.0022590	0.03314	. |*.
10	-0.0035152	-.05156	.*| .

"." marca dos errores estándar

Comprobación de la correlación cruzada entre series

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Correlaciones cruzadas					
5	117.75	6	<.0001	-0.005	0.057	-0.025	-0.282	-0.330	-0.453

Se ha eliminado el ruido blanco de ambas variables con el siguiente filtro:

Filtro para eliminar ruidos blancos

Factores autoregresivos

Factor 1: 1 - 1.97607 B**(1) + 1.37499 B**(2) - 0.34336 B**(3)

A continuación, se presenta la ***estimación del modelo de la función de transferencia sin estructura en el término de ruido***. La función PLOT imprime las funciones de autocorrelación residuales para este modelo se identifica un AR(2) para el ruido.

Estimación por mínimos cuadrados condicional

Parámetro	Estimación	Error estándar	Valor t	Aprox Pr > |t|	Retardo	Variable	Shift
MU	53.32237	0.04932	1081.24	<.0001	0	y	0
NUM1	-0.62868	0.25385	-2.48	0.0138	0	x	3
NUM1,1	0.47258	0.62253	0.76	0.4484	1	x	3
NUM1,2	0.73660	0.81006	0.91	0.3640	2	x	3
DEN1,1	0.15411	0.90483	0.17	0.8649	1	x	3
DEN1,2	0.27774	0.57345	0.48	0.6285	2	x	3

Constante Estimación	53.32237
Varianza Estimación	0.704241
Error estand Estimación	0.839191
AIC	729.7249
SBC	751.7648
Número de residuales	291

* AIC y SBC no incluyen determinante de la log.

Correlations of Parameter Estimates

Parámetro de la variable		y MU	x NUM1	x NUM1,1	x NUM1,2	x DEN1,1	x DEN1,2
y	MU	1.000	0.013	0.002	-0.002	0.004	-0.006
x	NUM1	0.013	1.000	0.755	-0.447	0.089	-0.065

x	NUM1,1	0.002	0.755	1.000	0.121	-0.538	0.565
x	NUM1,2	-0.002	-0.447	0.121	1.000	-0.892	0.870
x	DEN1,1	0.004	0.089	-0.538	-0.892	1.000	-0.998
x	DEN1,2	-0.006	-0.065	0.565	0.870	-0.998	1.000

Autocorrelation Check of Residuals

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	496.45	6	<.0001	0.893	0.711	0.502	0.312	0.167	0.064
12	498.58	12	<.0001	-0.003	-0.040	-0.054	-0.040	-0.022	-0.021
18	539.38	18	<.0001	-0.045	-0.083	-0.131	-0.170	-0.196	-0.195
24	561.87	24	<.0001	-0.163	-0.102	-0.026	0.047	0.106	0.142
30	585.90	30	<.0001	0.158	0.156	0.131	0.081	0.013	-0.037
36	592.42	36	<.0001	-0.048	-0.018	0.038	0.070	0.079	0.067
42	593.44	42	<.0001	0.042	0.025	0.013	0.004	0.006	0.019
48	601.94	48	<.0001	0.043	0.068	0.084	0.082	0.061	0.023

Autocorrelation Plot of Residuals

Retardo	Covarianza	Correlación	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Error Std.
0	0.704241	1.00000	|**********|	0
1	0.628846	0.89294	| . |**********|	0.058621
2	0.500490	0.71068	| . |**********|	0.094427
3	0.353404	0.50182	| . |**********|	0.111300
4	0.219895	0.31224	| . |******|	0.118821
5	0.117330	0.16660	| . |*** . |	0.121608
6	0.044967	0.06385	| . |* . |	0.122390
7	-0.0023551	-.00334	| . | . |	0.122504
8	-0.028030	-.03980	| . *| . |	0.122505
9	-0.037891	-.05380	| . *| . |	0.122549
10	-0.028378	-.04030	| . *| . |	0.122630

"." marca dos errores estándar

Autocorrelaciones inversas

Retardo	Correlación	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.57346	|**********| . |
2	0.02264	| . | . |
3	0.03631	| . |*. |
4	0.03941	| . |*. |
5	-0.01256	| . | . |
6	-0.01618	| . | . |
7	0.02680	| . |*. |
8	-0.05895	| .*| . |
9	0.07043	| . |*. |
10	-0.02987	| .*| . |

Autocorrelaciones parciales

Retardo	Correlación	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.89294	| . |**********| |
2	-0.42765	|*********| . |
3	-0.13463	|***| . |
4	0.02199	| . | . |
5	0.03891	| . |*. |
6	-0.02219	| . | . |
7	-0.02249	| . | . |
8	0.01538	| . | . |
9	0.00634	| . | . |
10	0.07737	| . |** |

Crosscorrelation Check of Residuals with Input x

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Correlaciones cruzadas					
5	0.48	2	0.7855	-0.009	-0.005	0.026	0.013	-0.017	-0.022
11	0.93	8	0.9986	-0.006	0.008	0.022	0.023	-0.017	-0.013
17	2.63	14	0.9996	0.012	0.035	0.037	0.039	-0.005	-0.040
23	19.19	20	0.5092	-0.076	-0.108	-0.122	-0.122	-0.094	-0.041
29	20.12	26	0.7857	-0.039	-0.013	0.010	-0.020	-0.031	-0.005
35	24.22	32	0.8363	-0.022	-0.031	-0.074	-0.036	0.014	0.076
41	30.66	38	0.7953	0.108	0.091	0.046	0.018	0.003	0.009
47	31.65	44	0.9180	0.008	-0.011	-0.040	-0.030	-0.002	0.028

Modelo para la variable y

Término independiente estimado 53.32237

Número de entrada 1

Variable de entrada x
Desplazamiento 3

Factores numeradores

Factor 1: -0.6287 - 0.47258 B**(1) - 0.7366 B**(2)

Factores denominadores

Factor 1: 1 - 0.15411 B**(1) - 0.27774 B**(2)

A continuación, ***ya podemos estimar el modelo de la función de transferencia completo*** con un estructura AR(2) en el ruido.

Estimación por mínimos cuadrados condicional

Parámetro	Estimación	Error estándar	Valor t	Aprox Pr > |t|	Retardo	Variable	Shift
MU	53.26307	0.11926	446.63	<.0001	0	y	0
AR1,1	1.53292	0.04754	32.25	<.0001	1	y	0
AR1,2	-0.63297	0.05006	-12.64	<.0001	2	y	0
NUM1	-0.53522	0.07482	-7.15	<.0001	0	x	3
NUM1,1	0.37602	0.10287	3.66	0.0003	1	x	3
NUM1,2	0.51894	0.10783	4.81	<.0001	2	x	3
DEN1,1	0.54842	0.03822	14.35	<.0001	1	x	3

Constante Estimación	5.329371
Varianza Estimación	0.058828
Error estand Estimación	0.242544
AIC	8.292811
SBC	34.00607
Número de residuales	291

* AIC y SBC no incluyen determinante de la log.

Correlations of Parameter Estimates

Parámetro de la variable		y MU	y AR1,1	y AR1,2	x NUM1	x NUM1,1	x NUM1,2	x DEN1,1
y	MU	1.000	-0.063	0.047	-0.008	-0.016	0.017	-0.049
y	AR1,1	-0.063	1.000	-0.927	-0.003	0.007	-0.002	0.015
y	AR1,2	0.047	-0.927	1.000	0.023	-0.005	0.005	-0.022
x	NUM1	-0.008	-0.003	0.023	1.000	0.713	-0.178	-0.013
x	NUM1,1	-0.016	0.007	-0.005	0.713	1.000	-0.467	-0.039
x	NUM1,2	0.017	-0.002	0.005	-0.178	-0.467	1.000	-0.720
x	DEN1,1	-0.049	0.015	-0.022	-0.013	-0.039	-0.720	1.000

Autocorrelation Check of Residuals

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Autocorrelaciones					
6	8.61	4	0.0717	0.024	0.055	-0.073	-0.054	-0.054	0.119
12	15.43	10	0.1172	0.032	0.028	-0.081	0.047	0.022	0.107
18	21.13	16	0.1734	-0.038	0.052	-0.093	-0.013	-0.073	-0.005
24	27.52	22	0.1922	-0.118	-0.002	-0.007	0.076	0.024	-0.004
30	36.94	28	0.1202	0.034	-0.021	0.020	0.094	-0.118	0.065
36	44.26	34	0.1119	-0.025	-0.057	0.113	0.022	0.030	0.065
42	45.62	40	0.2500	-0.017	-0.036	-0.029	-0.013	-0.033	0.017
48	48.60	46	0.3689	0.024	0.069	0.024	0.017	0.022	-0.044

Crosscorrelation Check of Residuals with Input x

Para retardo	Chi-cuadrado	DF	Pr > ChiSq	Correlaciones cruzadas					
5	0.93	3	0.8191	0.008	0.004	0.010	0.008	-0.045	0.030
11	6.60	9	0.6784	0.075	-0.024	-0.019	-0.026	-0.111	0.013
17	13.86	15	0.5365	0.050	0.043	0.014	0.014	-0.141	-0.028
23	18.55	21	0.6142	-0.074	-0.078	0.023	-0.016	0.021	0.060
29	27.99	27	0.4113	-0.071	-0.001	0.038	-0.156	0.031	0.035
35	35.18	33	0.3654	-0.014	0.015	-0.039	0.028	0.046	0.142
41	37.15	39	0.5544	0.031	-0.029	-0.070	-0.006	0.012	-0.004
47	42.42	45	0.5818	0.036	-0.038	-0.053	0.107	0.029	0.021

Modelo para la variable y

Término independiente estimado 53.26307

Factores autoregresivos

Factor 1: 1 - 1.53292 B**(1) + 0.63297 B**(2)

Número de entrada 1

Variable de entrada x
Desplazamiento 3

Factores numeradores

Factor 1: -0.5352 - 0.37602 B**(1) - 0.51894 B**(2)

Factores denominadores

Factor 1: 1 - 0.54842 B**(1)

El modelo de función de transferencia identificado, estimado y diagnosticado correctamente es el siguiente:

$$\boxed{(1-0{,}53292B+0{,}63297B^2)Y_t=-199{,}622+\frac{-0.5352-0{,}37602B-0{,}51894B^2}{(1-0{,}54842B)}X_t+a_t}$$

CAPÍTULO 8

MODELOS DEL ANÁLISIS DE LA VARIANZA Y LA COVARIANZA. MODELO LINEAL GENERAL GLM Y MODELOS MIXTOS

MODELOS DEL ANÁLISIS DE LA VARIANZA Y LA COVARIANZA

Los *modelos del análisis de la varianza y la covarianza simples* tienen como característica fundamental que la variable dependiente es cuantitativa y las variables independientes son cualitativas o mezcla de cualitativas con cuantitativas.

El *análisis de la varianza simple* es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica (cuantitativa) y varias variables independientes (o exógenas) no métricas (cualitativas). Los valores no métricos (categorías) de las variables independientes determinarán una serie de grupos en la variable dependiente. El modelo ANOVA mide la significación estadística de las diferencias entre las medias de los grupos determinados en la variable dependiente por los valores (categorías) de las variables independientes no métricas (cualitativas).

La expresión funcional del modelo del análisis de la varianza simple ANOVA con variable dependiente métrica (y) y variables independientes no métricas $(x_1, x_2, \cdots, x_n)$ es:

$$y = F(x_1, x_2, \cdots, x_n)$$

El *análisis de la covarianza simple* es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas), parte de las cuales son no métricas, siendo la otra parte métricas (*covariables*).

La expresión funcional del modelo del análisis de la covarianza simple ANCOVA es la siguiente:

$$y = F(x_1, x_2, \cdots, x_n)$$

La variable dependiente y es métrica y las variables independientes son algunas métricas y otras no métricas.

Los *modelos del análisis de la varianza y la covarianza múltiples* tienen como característica fundamental la presencia de varias variables dependientes cuantitativas y varias variables independientes cualitativas o mezcla de cualitativas con cuantitativas.

El *análisis de la varianza múltiple* es una técnica estadística utilizada para analizar la relación entre varias variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) no métricas. El objetivo esencial de los modelos del análisis de la varianza múltiple es contrastar si los valores no métricos de las variables independientes determinarán la igualdad de vectores de medias de una serie de grupos determinados por ellos en las variables dependientes. De modo que el modelo MANOVA mide la significación estadística de las diferencias entre los vectores de medias de los grupos determinados en las variables dependientes por los valores de las variables independientes.

La expresión funcional del modelo del análisis de la varianza múltiple MANOVA es la siguiente:

$$G(y_1, y_2, \cdots, y_m) = F(x_1, x_2, \cdots, x_n)$$

Las variables dependientes $(y_1, y_2, \cdots, y_m)$ son métricas y las variables independientes $(x_1, x_2, \cdots, x_n)$ son no métricas.

El *análisis de la covarianza múltiple* es una técnica estadística utilizada para analizar la relación entre varias variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) mezcla de variables métricas y no métricas.

La expresión funcional del modelo del análisis de la covarianza múltiple MANCOVA es la siguiente:

$$G(y_1, y_2, \cdots, y_m) = F(x_1, x_2, \cdots, x_n)$$

Las variables dependientes $(y_1, y_2, \cdots, y_m)$ son métricas y las variables independientes $(x_1, x_2, \cdots, x_n)$ son una parte métricas y otra parte no métricas.

En el análisis de la covarianza, tanto simple como múltiple, las variables métricas independientes (*covariables*) tienen como objetivo eliminar determinados efectos que puedan sesgar los resultados incrementando la varianza dentro de los grupos.

En el análisis de la covarianza se suele comenzar eliminando, mediante una regresión lineal, la variación experimentada por las variables dependientes producida por la covariable o covariables de efectos indeseados, para continuar con un análisis ANOVA o MANOVA sobre las variables dependientes ajustadas (residuos de la regresión anterior).

MODELOS ANOVA DE LA VARIANZA SIMPLE

Modelo ANOVA con un solo factor. Efectos fijos y aleatorios

El *análisis de la varianza simple* (ANOVA) con un solo factor se presenta cuando analizamos la relación entre una variable dependiente métrica (endógena o *variable respuesta*) y una variable independiente no métrica (*factor* o exógena) estudiada en sus distintos niveles o grupos (los G valores que puede tomar o *tratamientos*).

En el caso del ***modelo con efectos aleatorios*** consideraremos de modo general que de la población total de niveles del factor (supuesta infinita o suficientemente grande como para considerarla infinita), los G niveles del factor que se utilizan en el experimento se han elegido aleatoriamente. En este caso, el modelo ANOVA de efectos aleatorios se formulará de la siguiente forma:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Una formulación equivalente, si consideramos que $\mu_i = \mu + \beta_i$, sería:

$$Y_{ij} = \mu + \beta_i + \varepsilon_{ij}$$

donde:

- μ es una constante.
- β_i, para $i = 1, \ldots, G$, son variables aleatorias independientemente distribuidas $N(0, \sigma^2{}_\beta)$.
- ε_{ij}, para $i = 1, \ldots, G$ y $j = 1, \ldots, n_i$, son v.a.i.i.d. $N(0, \sigma^2)$.
- β_i y ε_{ij}, para $i = 1, \ldots, G$ y $j = 1, \ldots, n_i$, son variables aleatorias independientes. (v.a.i.i.d.=variables aleatorias independientes idénticamente distribuidas)

Para este modelo se verifica que $E[Y_{ij}] = \mu$, y la varianza de Y_{ij}, denotada por σ_Y^2, será $V[Y_{ij}] = \sigma_Y{}^2 = \sigma^2{}_\beta + \sigma^2$, donde $\sigma^2{}_\beta$ y σ^2 se llaman *componentes de la varianza*, razón por la cual el modelo se denomina *modelo de componentes de la varianza*.

Una vez estimado el modelo, será necesario comprobar mediante distintos tests si las hipótesis básicas del mismo no están en contradicción con los datos observados. Para contrastar la normalidad de los errores experimentales ε_{ij} (hipótesis de normalidad), suele usarse el test W de Shapiro y Wilk, un contraste de la *Chi-cuadrado* o el test de Kolmogorov-Smirnov. Para contrastar la igualdad de varianzas de los ε_{ij} (hipótesis de homoscedasticidad), suele usarse el test de Barlett, el test Q de Cochran, el test de Siegel-Tukey y el test de Hartley. Para contrastar la independencia de las observaciones, o no correlación de los residuos (hipótesis de no autocorrelación), suelen utilizarse el coeficiente de correlación serial o el test de rachas.

Ya sabemos que es interesante contrastar si es aceptable la hipótesis de que las medias de todos los grupos de observaciones obtenidas al repetir el experimento para cada nivel de factor son idénticas ($u_1 = u_2 = ... = u_G = u$). Si los contrastes diesen como resultado que esta hipótesis es cierta, la pertenencia a un grupo o a otro sería irrelevante, y podríamos considerar todas las observaciones como una muestra de una única población. Un enfoque alternativo de esta hipótesis, que conduce al mismo resultado, es considerar los grupos idénticos si las diferencias entre sus medias son pequeñas.

Se pueden construir *intervalos de confianza para las diferencias entre medias* de distintos grupos ($u_i - u_j$), con el fin de estimar si existen diferencias entre ellos. Como norma general, si el intervalo contiene al cero se suele aceptar la hipótesis de medias iguales para los grupos. También se pueden construir intervalos de confianza para la varianza del error experimental y para cocientes de varianzas.

En general, cuando estudiamos el comportamiento de los niveles de un factor, no se persigue como única finalidad del análisis de la varianza saber si globalmente los distintos niveles de factor son significativamente distintos entre sí en su efecto sobre la variable respuesta (aspecto de un evidente interés), sino que, lógicamente, estaremos interesados en conocer, una vez contrastado que las diferencias son significativas, qué niveles producen un efecto superior al de otros sobre la variable respuesta.

Para ello existen diferentes contrastes que efectúan comparaciones múltiples entre las G medias o combinaciones lineales de ellas (test de recorrido múltiple de Duncan, test de Bonferroni, test SNK de Student-Newman-Keuls, test HSD de Tukey, test de la diferencia mínima significativa, test de Scheffe y test de Tukey de comparaciones múltiples).

También se puede considerar el contraste:

$$H_0: \sigma^2_\beta = 0$$
$$H_1: \sigma^2_\beta > 0$$

Hemos supuesto hasta aquí que los niveles estudiados de factor son una muestra de la población de niveles que se supone infinita, es decir, los factores son aleatorios. Se dice entonces que estamos ante un *modelo de efectos aleatorios*.

Sin embargo, un modelo de análisis de la varianza es un ***modelo de efectos fijos*** cuando los resultados obtenidos sólo son válidos para esos determinados niveles de factor estudiados en el momento actual (factores constantes), pudiendo ser diferente lo que ocurra a otros niveles de factor.

El cuadro del análisis de la varianza es el mismo para efectos fijos y aleatorios. La única diferencia es que en efectos aleatorios hay que considerar las componentes de la varianza (con la presencia de σ_β^2) y en los efectos fijos no.

Modelo ANOVA con dos factores: efectos fijos, aleatorios y mixtos

Supongamos ahora que tenemos dos factores de modo que cada uno de los dos conjuntos de niveles de los factores se puede considerar una muestra de una población suficientemente grande sobre la que se van a realizar estudios. En este caso, se dice que estamos en presencia del ***modelo ANOVA bifactorial general con efectos aleatorios***, que se formula de la siguiente forma:

$$Y_{ijl} = \mu + \beta_i + \delta_j + (\beta\delta)_{ij} + \varepsilon_{ijl} \qquad \beta\delta = \textit{interacción entre los dos factores}$$

donde, para $i = 1,...,h$; $j = 1,...,k$; $l = 1,...,t$; se verifica que μ es una constante, β_i son v.a. independientes distribuidas $N(0, \sigma_\beta^2)$, δ_j son v.a. independientes distribuidas $N(0, \sigma_\delta^2)$, $(\beta\delta)_{ij}$ son v.a. independientes distribuidas $N(0, \sigma_{\beta\delta}^2)$, ε_{ijl} son v.a. independientes distribuidas $N(0, \sigma^2)$, y β_i, δ_j, $(\beta\delta)_{ij}$ y ε_{ijl} son v.a. independientes dos a dos.

Para este modelo se verifica que $E[Y_{ijl}] = \mu$, y la varianza de Y_{ijl}, notada por σ_y^2, viene dada por $V[Y_{ijl}] = \sigma_y^2 = \sigma_\beta^2 + \sigma_\delta^2 + \sigma^2$.

Para estimar las componentes de la varianza, σ^2, σ_β^2, σ_δ^2 y $\sigma_{\beta\delta}^2$, se utilizan los estimadores CME, $(CMA - CMAB)/k.t$, $(CMB - CMAB)/h.t$ y $(CMAB - CME)/t$ respectivamente.

Lo visto hasta aquí corresponde a un modelo bifactorial con los dos factores aleatorios. Pero en un modelo factorial de dos factores, pueden ser los dos fijos (*modelo bifactorial de efectos fijos*), los dos aleatorios (*modelo bifactorial de efectos aleatorios*), o uno aleatorio y otro fijo (*modelo bifactorial de efectos mixtos*).

Supongamos un ***modelo ANOVA bifactorial de efectos mixtos***. Si el factor A tiene niveles fijos y el factor B tiene niveles aleatorios, los efectos βi, son constantes, los efectos δ_j son v.a., y los efectos de interacción $(\beta\delta)_{ij}$ también son v.a. al serlo los δ_j. Suponiendo tamaños muestrales iguales para cada tratamiento, se tiene que el modelo bifactorial general con efectos mixtos se formula de la siguiente forma:

$$Y_{ijl} = \mu + \beta_i + \delta_j + (\beta\delta)_{ij} + \varepsilon_{ijl}$$

donde se verifica que μ es una constante, β_i son constantes tales que $\Sigma\beta_i = 0$, δ_j para $j = 1,...,k$ son v.a. independientes distribuidas $N(0, \sigma_{\beta\delta}^2)$, $(\beta\delta)_{ij}$ para $i = 1,...,h$, son v.a.i.i.d. $N(0, \sigma_{\beta\delta}^2(h-1/h))$ sujetas a las restricciones $\Sigma(\beta\delta)_{ij} = 0$ para todo $j = 1,...,k$, ε_{ijl} para $i=1,...,h$, $j=1,...,k$ y $l=1,...,t$ son v.a. independientes distribuidas $N(0, \sigma^2)$, y δ_j, $(\beta\delta)_{ijl}$ y ε_{ijl} son v.a. independientes dos a dos.

Sólo resta por analizar el ***modelo ANOVA bifactorial de efectos fijos***. Para un modelo factorial de dos factores A y B de efectos fijos tendremos la expresión general:

$$X_{ijk}=u+A_i+B_j+AB_{ij}+E_{ijk} \quad i=1,...,t,\ j=1,...,r,\ k=1,...,n_{ij} \quad (u = \text{constante})$$

Los términos A_i y B_j representan los efectos de los factores A y B (*efectos principales*), y son constantes sujetas a las restricciones:

$$\sum_{i=1}^{t} A_i = \sum_{j=1}^{r} B_j = 0$$

Los términos AB_{ij} representan el *efecto de la interacción* entre los factores A y B, y son constantes sujetas a las restricciones:

$$\sum_{i=1}^{t} (AB)_{ij} = \sum_{j=1}^{r} (AB)_{ij} = 0$$

El término E_{ijk} representa el error experimental, que corresponderá a una variable aleatoria normal de media cero y varianza σ^2 constante para cada k (las variables E_{ijk} han de ser independientes).

El desarrollo del modelo se ha realizado bajo la hipótesis del mismo número de observaciones, s, para cada tratamiento o combinación de niveles de los factores. Cuando no se cumple dicha hipótesis, el análisis de la varianza para un estudio de dos factores se hace más complejo, y ya no son válidas las fórmulas de descomposición de suma de cuadrados, aunque se mantiene la filosofía de la descomposición de la suma de cuadrados total y de los grados de libertad. A partir de las nuevas expresiones de sumas de cuadrados, se consideran los cuadrados medios y el cociente correspondiente entre cuadrados medios, de manera análoga. Una forma sencilla de abordar este problema y obtener las sumas de cuadrados apropiadas para realizar los contrastes de hipótesis sobre los efectos de interacción de los factores y sobre los efectos principales de los factores, es considerar el análisis de la varianza desde la perspectiva del análisis de regresión.

Los intervalos de confianza para combinaciones lineales de tratamientos se llevan a cabo de la misma forma que en el modelo de un solo factor, pero teniendo en cuenta que si las interacciones son significativamente distintas de cero, se utilizará *CMAB* en lugar de *CME* en las fórmulas.

Modelo ANOVA con tres factores

Para un modelo factorial de tres factores A, B y C, tendríamos la expresión general:

$$X_{ijkl} = u+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}*(\alpha\beta\gamma)_{ijk}+\delta_{ijkl} \quad i=1,...,t,\ j=1,...,r,\ k=1,...,s,\ l=1,...,n_{ijk}$$

Los términos α_i, β_j y γ_k representan los efectos de los factores A, B y C (efectos principales). El término AB_{ij} representa el efecto de la interacción entre los factores A y B.

El término $(\alpha\gamma)_{ij}$ representa el efecto de la interacción entre los factores A y C. El término $(\beta\gamma)_{ij}$ representa el efecto de la interacción entre los factores B y C. El término $(\alpha\beta\gamma)_{ijk}$ representa la interacción triple entre los factores A, B y C. El término δ_{ijkl} representa el error experimental, que corresponderá a una variable aleatoria normal de media cero y varianza constante para cada l. Las variables δ_{ijkl} han de ser independientes. El modelo también puede considerarse con término constante.

En un modelo factorial de tres factores, pueden ser los tres fijos, los tres aleatorios, uno aleatorio y dos fijos, o dos aleatorios y el otro fijo. En un modelo multifactorial los niveles de cada factor (tratamientos) suelen estar combinados con todos los niveles de los restantes factores. En el caso de que ciertos niveles de determinados factores estén ligados solamente a ciertos niveles de otros, estamos ante un *diseño jerárquico*.

En un diseño jerárquico, los niveles de cada factor están incluidos en los niveles de otro factor, estableciéndose así una jerarquía de dependencias entre los distintos niveles de los diferentes factores. Un modelo jerárquico es *anidado* cuando cada nivel de un factor se corresponde sólo con un nivel de otro factor. En este tipo de modelos no existen interacciones, ya que esto sólo es posible cuando todos los niveles de un determinado factor se cruzan con todos los niveles de los demás factores. Un modelo jerárquico es *cruzado* cuando todos los niveles de un factor aparecen en todos los niveles de resto de los factores.

MODELOS ANCOVA DE LA COVARIANZA SIMPLE

Si ampliamos el análisis de la varianza suponiendo que influyen en la variable respuesta (variable independiente), además de los factores, una o varias variables cuantitativas, se aplicará un análisis de la covarianza ANCOVA para explicar correctamente dicha variable respuesta. Estas variables cuantitativas se denominan ***covariantes*** o ***variables concomitantes***.

De una manera muy general puede considerarse que el análisis de la covarianza reúne las técnicas del análisis de la varianza y del análisis de la regresión. La diferencia entre análisis de la varianza y análisis de la regresión radica en la forma de tratar las variables independientes (factores).

En el análisis de la regresión todos los factores son cuantitativos y se tratan cuantitativamente. En el análisis de la varianza los factores suelen ser cualitativos, pero si alguno es cuantitativo, se trata cualitativamente. En el análisis de la covarianza, por ser una mezcla de ambos análisis, unos factores se tratan cualitativamente (los factores del análisis de la varianza) y otros cuantitativamente (los covariantes).

Modelo con un factor y un covariante

El modelo de análisis de la covarianza más simple que se puede considerar es el que tiene un factor y un covariante, y será de la forma:

$$Y_{ij} = u + A_i + \beta X_{ij} + E_{ij} \quad i = 1, \ldots, t \quad j = 1, \ldots, n_i$$

donde A_i es el factor fijo y X_{ij} es el covariante. Notemos que X_{ij} no es una variable aleatoria. El error E_{ij} sí es una variable aleatoria, con las hipótesis de normalidad, homocedasticidad, independencia y esperanza matemática nula.

Las variables E_{ij} son normales $N(0,\sigma)$ e independientes, y por ser el factor A fijo, sus distintos niveles verificarán la condición $\sum_{i=1}^{t} n_i A_i = 0$.

Si $n_i = n$ para todo $i = 1, \ldots, t$, el modelo es equilibrado.

Modelo con dos factores y un covariante

Para un modelo de dos factores y un covariante tendremos la expresión:

$$Y_{ij} = u + A_i + B_j + \beta X_{ij} + E_{ij} \quad i = 1, \ldots, t, \quad j = 1, \ldots, n_i$$

donde A_i y B_j son los factores fijos y X_{ij} es el covariante. Obsérvese que X_{ij} no es una variable aleatoria. El error E_{ij} sí es una variable aleatoria, con las hipótesis de normalidad, homocedasticidad, independencia y esperanza matemática nula.

Las variables E_{ij} son normales $N(0,\sigma)$ e independientes, y por ser los factores A y B fijos, sus distintos niveles verificarán la condición:

$$\sum_{i=1}^{t} A_i = \sum_{j=1}^{n} B_j = 0$$

Si consideramos $n_i = n$ para todo $i = 1, \ldots, t$, el modelo es equilibrado.

Modelo con dos factores y dos covariantes

Para un modelo de dos factores y dos covariantes tendremos la expresión:

$$Y_{ij} = u + A_i + B_j + \gamma X_{ij} + \delta W_{ij} + E_{ij} \quad i = 1, \ldots, t, \quad j = 1, \ldots, n_i$$

donde A_i y B_j son los factores fijos y X_{ij} y W_{ij} son los covariantes. Observemos que X_{ij} y W_{ij} no son variables aleatorias. El error E_{ij} sí es una variable aleatoria, con las hipótesis de normalidad, homocedasticidad, independencia y esperanza matemática nula.

Las variables E_{ij} son normales $N(0,\sigma)$ e independientes, y por ser los factores A y B fijos, sus distintos niveles verificarán la condición:

$$\sum_{i=1}^{t} A_i = \sum_{j=1}^{n} B_j = 0.$$

Si consideramos $n_i = n$ para todo $i = 1, \ldots, t$, el modelo es equilibrado.

ANÁLISIS MULTIVARIANTE DE LA VARIANZA (MANOVA)

El ***análisis multivariante de la varianza*** o ***de la varianza múltiple MANOVA*** es una técnica estadística utilizada para analizar la relación entre varias variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) no métricas. El objetivo esencial de los modelos del análisis de la varianza múltiple es contrastar si los valores no métricos de las variables independientes determinarán la igualdad de vectores de medias de una serie de grupos determinados por ellos en las variables dependientes. De modo que el modelo MANOVA mide la significación estadística de las diferencias entre los vectores de medias de los grupos determinados en las variables dependientes por los valores de las variables independientes.

Análisis multivariante de la varianza con un factor

El modelo subyacente en el análisis multivariante de la varianza con un factor y K variables dependientes, viene dado por la siguiente expresión:

$$\begin{bmatrix} Y_{1G} \\ Y_{2G} \\ \cdots \\ Y_{Kg} \end{bmatrix} = \begin{bmatrix} \mu_{1G} \\ \mu_{2G} \\ \cdots \\ \mu_{Kg} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1G} \\ \varepsilon_{2G} \\ \cdots \\ \varepsilon_{Kg} \end{bmatrix} \qquad g = 1,2,\cdots,G$$

que en forma matricial puede escribirse como: $y_g = \mu_g + \varepsilon_g$

En el modelo se han considerado G grupos del factor analizado y el vector y_g es igual a la media teórica del grupo (μ_g) más una variable aleatoria (ε_g). El definir que el valor medio teórico de y_g es μ_g implica que la media teórica de ε_g es igual a 0. Se requieren dos subíndices para referirse a una variable determinada. Así, el primer subíndice de Y_{2g} se refiere a la variable 2 de entre K variables y el segundo subíndice hace referencia al grupo g de entre G grupos.

Las hipótesis estadísticas, que se adoptan para la aplicación del análisis multivariante de la varianza se refieren tanto a la población como al proceso de obtención de la muestra. Las hipótesis sobre la población son que $y_g \rightarrow N(\mu_g,\Sigma)$ siendo Σ la matriz de covarianzas de todas las poblaciones (hipótesis de homoscedasticidad) y cada una de las poblaciones tiene una distribución normal multivariante. La hipótesis sobre el proceso de obtención de la muestra facilita la realización del proceso de inferencia a partir de la información disponible, y en nuestro caso se resume en que se supone que se ha extraído una muestra aleatoria multivariante independiente en cada una de las G poblaciones.

En el modelo de un factor, las hipótesis nula y alternativa a contrastar son las siguientes:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_G$
H_1 : No todas μ_g son iguales

Análisis multivariante de la varianza con dos factores

El modelo subyacente en el ***análisis multivariante de la varianza con dos factores*** A (g niveles) y B (j niveles), y con K variables dependientes, viene dado por la siguiente expresión:

$$\begin{bmatrix} Y_{1gj} \\ Y_{2gj} \\ \cdots \\ Y_{Kgj} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_K \end{bmatrix} + \begin{bmatrix} \alpha_{1g} \\ \alpha_{2g} \\ \cdots \\ \alpha_{Kg} \end{bmatrix} + \begin{bmatrix} \alpha_{1g} \\ \alpha_{2g} \\ \cdots \\ \alpha_{Kg} \end{bmatrix} + \begin{bmatrix} \beta_{1j} \\ \beta_{2j} \\ \cdots \\ \beta_{Kj} \end{bmatrix} + \begin{bmatrix} (\alpha\beta)_{1gj} \\ (\alpha\beta)_{2gj} \\ \cdots \\ (\alpha\beta)_{2gj} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1gj} \\ \varepsilon_{2gj} \\ \cdots \\ \varepsilon_{Kgj} \end{bmatrix} \qquad \begin{cases} g = 1,2,\cdots,G \\ j = 1,2,\cdots,J \end{cases}$$

que en forma matricial puede escribirse como: $y_{gj} = \mu + \alpha_g + \beta_j + (\alpha\beta)_{gj} + \varepsilon_{gj}$

En el modelo anterior μ es la media general, α_g es el efecto diferencial del nivel g del factor A, β_j es el efecto del nivel j del factor B y $(\alpha\beta)_{gj}$ es el efecto interacción de los niveles g y j.

En el análisis multivariante de la varianza con dos factores se mantienen las hipótesis de poblaciones con distribución normal multivariante e igual matriz de covarianzas para cada combinación de niveles de los dos factores. Adicionalmente, se pueden formular hipótesis para realizar contrastes sobre el factor A, sobre el B y sobre su interacción.

La descomposición de la matriz T (denominada también SCPC total) en el análisis multivariante de la varianza es la siguiente:

$$T = F_A + F_B + F_{AB} + W$$

El estadístico lambda de Wilks en el caso de dos factores permite contrastar tres hipótesis relativas al factor A, al factor B y a la interacción entre ambos (cuyas matrices son F_A, F_B y F_{AB} respectivamente). El estadístico tiene la siguiente expresión:

$$\Lambda = \frac{|W|}{|F_H + W|} \qquad H = A, B, AB$$

ANÁLISIS MULTIVARIANTE DE LA COVARIANZA (MANCOVA)

Ya sabemos que el análisis de la covarianza múltiple es una técnica estadística utilizada para analizar la relación entre varias variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) mezcla de variables métricas y no métricas. Se trata entonces de una generalización del modelo del análisis multivariante de la varianza con la inclusión de variables independientes cuantitativas denominadas covariables. En el ***modelo MANCOVA*** las variables dependientes son métricas y las variables independientes son una parte métricas y otra parte no métricas.

MODELO LINEAL GENERAL (GLM)

El ***modelo de regresión Múltiple Lineal General (GLM)*** es el modelo más general posible de regresión lineal, incluyendo el modelo de regresión lineal múltiple con variables cuantitativas y los modelos de regresión múltiple con variables cualitativas y cuantitativas a la vez, por lo que incluirá todos los modelos del análisis de la varianza y de la covarianza. Este modelo es implementado por la mayoría del software especializado actual para resolver cualquier tipo de problema del análisis de la regresión, análisis de la varianza y análisis de la covarianza.

MODELOS LINEALES MIXTOS

Los ***modelos lineales mixtos*** expanden el modelo lineal general de modo que los datos puedan presentar variabilidad correlacionada y no constante. El modelo lineal mixto proporciona, por tanto, la flexibilidad necesaria para modelizar no sólo las medias sino también las varianzas y covarianzas de los datos. A partir de los modelos lineales mixtos se pueden estudiar los modelos multinivel, los modelos con datos de panel, los modelos lineales jerárquicos y los modelos con coeficientes aleatorios.

Los modelos lineales mixtos son una extensión del modelo lineal general que no requiere el cumplimiento de las hipótesis de independencia ni homoscedasticidad. Por tanto son adecuados para realizar estimaciones de modelos de panel. Por otro lado, el modelo lineal general permite la especificación de efectos aleatorios, pero realiza las estimaciones como si se tratase de efectos fijos. Las componentes de la varianza estimadas son calculadas a partir de cuadrados medios esperados, mientras que las estimaciones calculadas en los modelos mixtos se obtienen por máxima verosimilitud y máxima verosimilitud restringida.

El modelo lineal general permite analizar diseños con medidas repetidas, sin embargo, los modelos mixtos permiten trabajar con una gama variada de estructuras de la covarianza, incluyendo casos con datos incompletos.

Los modelos mixtos producen estadísticos (tamaños de las muestras, medias y desviaciones típicas de la variable dependiente y las covariables para cada combinación de niveles de los factores), información de los niveles del factor (valores ordenados de los niveles de cada factor y las frecuencias correspondientes), estimaciones de los parámetros, intervalos de confianza para los efectos fijos y las pruebas de Wald y los intervalos de confianza para los parámetros de las matrices de covarianzas.

En los modelos lineales mixtos la variable dependiente debe ser cuantitativa. Los factores deben ser categóricos y pueden tener valores numéricos o valores de cadena. Las covariables y la variable de ponderación deben ser cuantitativas. Las variables de sujetos y repetidas pueden ser de cualquier tipo.

Además se asume que la variable dependiente está relacionada linealmente con los factores fijos, los factores aleatorios y las covariables. Los efectos fijos modelan la media de la variable dependiente. Los efectos aleatorios modelan la estructura de las covarianzas de la variable dependiente. Los efectos aleatorios múltiples se consideran independientes entre sí y se calculan por separado las matrices de covarianzas de cada uno de ellos, sin embargo, se puede establecer una correlación entre los términos del modelo especificados para el mismo efecto aleatorio. Las medidas repetidas modelan la estructura de las covarianzas de los residuos. Se asume además que la variable dependiente procede de una distribución normal.

Es conveniente realizar un análisis exploratorio de los datos antes de realizar un análisis. Si no existe una variabilidad correlacionada o no constante, puede utilizarse el modelo lineal univariante o en medidas repetidas. Alternativamente, puede utilizarse análisis de componentes de la varianza en caso de que los efectos aleatorios tengan una estructura de covarianzas en los componentes de la varianza y no haya medidas repetidas.

SPSS Y LOS MODELOS ANOVA Y ANCOVA UNIVARIANTES DE UNO Y VARIOS FACTORES

SPSS incorpora el procedimiento ***MLG Univariante*** (o ***GLM Univariante***) que proporciona un análisis de regresión y un análisis de varianza para una variable dependiente cuantitativa y uno o más factores o variables cualitativas independientes. Las categorías de las variables de factor dividen la población en grupos. Con el procedimiento MLG (modelo lineal general) se pueden contrastar hipótesis nulas sobre los efectos de las variables independientes en las medias de varias agrupaciones de una única variable dependiente. Se pueden investigar las interacciones entre los factores así como los efectos de los factores individuales, algunos de los cuales pueden ser aleatorios. Además, se pueden incluir los efectos de las covariables y las interacciones de covariables con los factores. Para el análisis de regresión, las variables independientes (predictoras) se especifican como covariables (cuantitativas).

Se pueden contrastar tanto los modelos equilibrados como los no equilibrados. Se considera que un diseño está equilibrado si cada casilla del modelo contiene el mismo número de casos. Además de contrastar hipótesis, MLG Univariante genera estimaciones de los parámetros. También se encuentran disponibles los contrastes *a priori* de uso más habitual para contrastar las hipótesis. Además, si una prueba F global ha mostrado cierta significación, pueden emplearse las pruebas *post hoc* para evaluar las diferencias entre las medias específicas.

Las medias marginales estimadas ofrecen estimaciones de valores de las medias pronosticadas para las casillas del modelo; los gráficos de perfil (gráficos de interacciones) de estas medias permiten observar fácilmente algunas de estas relaciones. En su archivo de datos puede guardar residuos, valores pronosticados, distancia de Cook y valores de influencia como variables nuevas para comprobar los supuestos. Ponderación MCP permite especificar una variable usada para aplicar a las observaciones una ponderación diferente en un análisis de mínimos cuadrados ponderados (MCP), por ejemplo para compensar la distinta precisión de las medidas.

En cuanto a estadísticos, se obtienen las pruebas de rango *post hoc* y las comparaciones múltiples: diferencia menos significativa (DMS), Bonferroni, Sidak, Scheffé, múltiples F de Ryan-Einot-Gabriel-Welsch (R-E-G-WF), rango múltiple de Ryan-Einot-Gabriel-Welsch, Student-Newman-Keuls (S-N-K), diferencia honestamente significativa de Tukey, b de Tukey, Duncan, $GT2$ de Hochberg, Gabriel, pruebas t de Waller Duncan, Dunnett (unilateral y bilateral), $T2$ de Tamhane, T3 de Dunnett, Games-Howell y C de Dunnett. Estadísticos descriptivos: medias observadas, desviaciones típicas y frecuencias de todas las variables dependientes en todas las casillas. Prueba de Levene para la homogeneidad de varianzas. En cuanto a gráficos se obtienen diagramas de dispersión por nivel, gráficos de residuos y gráficos de perfil (interacción).

Para realizar un análisis MLG Univariante, elija en los menús *Analizar* → *Modelo lineal general* → *Univariante* (Figura 8-1), seleccione una variable dependiente, seleccione variables para *Factores fijos, Factores aleatorios* y *Covariables*, en función de los datos (Figura 8-2). En caso de existencia de heteroscedasticidad, para especificar una variable de ponderación, utilice *Ponderación MCP*. El botón *Especificar modelo* (Figura 8-3) permite definir un modelo factorial completo que contiene todos los efectos principales del factor, todos los efectos principales de las covariables y todas las interacciones factor por factor. No contiene interacciones de covariable. Seleccione *Personalizado* para especificar sólo un subconjunto de interacciones o para especificar interacciones factor por covariable. Indique todos los términos que desee incluir en el modelo. Como ejemplo, en el fichero *coches* usamos un modelo ANCOVA de variable dependiente *potencia*, con factor fijo *origen*, con factor aleatorio *cilindr* y con covariables *motor, peso* y *aceleración*.

El botón *Contrastes* (Figura 8-4) permite definir los contrastes *ad hoc* de las diferencias entre los niveles de cada factor. Puede especificar un tipo de contraste para cada factor en el modelo (en un modelo de medidas repetidas, para cada factor inter-sujeto). Los contrastes representan las combinaciones lineales de los parámetros y puede elegirse entre *Desviación* (todos los niveles de factor salvo el último se comparan con la media total), *Simple* (todos los niveles de factor salvo el último se comparan con la última categoría), *Diferencia* (cada nivel, excepto el primero, se compara con la media de los anteriores), *Helmert* (cada nivel, excepto el último, se compara con la media de las categorías posteriores), *Repetido* (cada nivel, excepto el primero, se compara con el nivel anterior), *Polinómico* (comparaciones de tendencia) y *Especial* (por sintaxis).

El botón *Gráficos* (Figura 8-5) permite definir los gráficos de perfil (gráficos de interacción) que sirven para comparar las medias marginales en el modelo. Un gráfico de perfil es un gráfico de líneas en el que cada punto indica la media marginal estimada de una variable dependiente (corregida respecto a las covariables) en un nivel de un factor. Los niveles de un segundo factor se pueden utilizar para generar líneas diferentes. Cada nivel en un tercer factor se puede utilizar para crear un gráfico diferente. Todos los factores fijos y aleatorios, si existen, están disponibles para los gráficos. Para los análisis multivariados, los gráficos de perfil se crean para cada variable dependiente. En un análisis de medidas repetidas, es posible utilizar tanto los factores inter-sujetos como los intra-sujetos en los gráficos de perfil. Las opciones *MLG - Multivariante y MLG - Medidas repetidas* sólo estarán disponibles si tiene instalada la opción *Modelos avanzados*. Un gráfico de perfil de un factor muestra si las medias marginales estimadas aumentan o disminuyen a través de los niveles. Para dos o más factores, las líneas paralelas indican que no existe interacción entre los factores, lo que significa que puede investigar los niveles de un único factor. Las líneas no paralelas indican una interacción.

El botón *Post hoc* ya fue explicado en el procedimiento anterior. El botón *Opciones* (Figura 8-6) permite seleccionar estadísticos adicionales. El botón *Guardar* permite guardar los valores pronosticados por el modelo, los residuos y las medidas relacionadas como variables nuevas en el *Editor de datos*.

Muchas de estas variables se pueden utilizar para examinar supuestos sobre los datos. Si desea almacenar los valores para utilizarlos en otra sesión de SPSS, guárdelos en el archivo de datos actual.

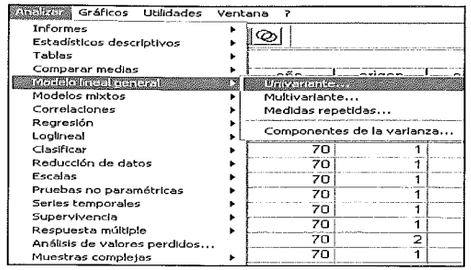


Figura 8-1

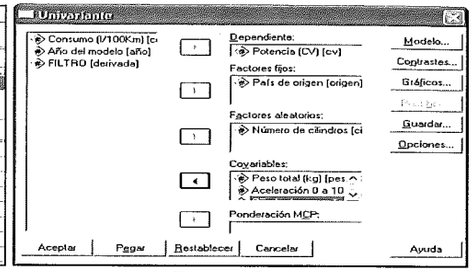


Figura 8-2

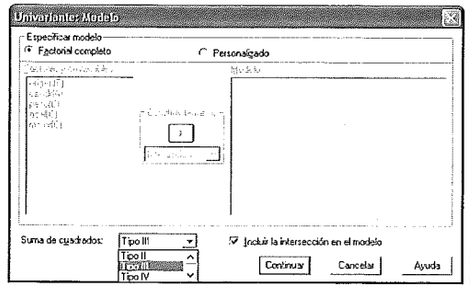


Figura 8-3

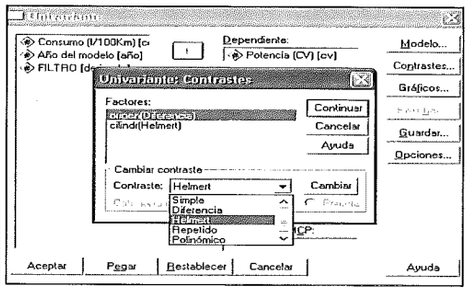


Figura 8-4

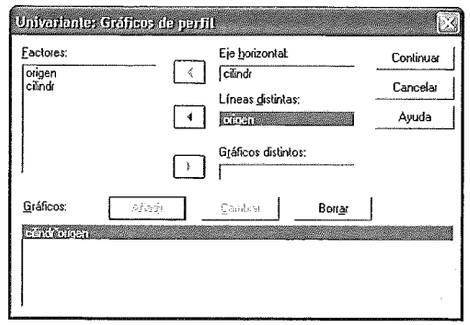


Figura 8-5

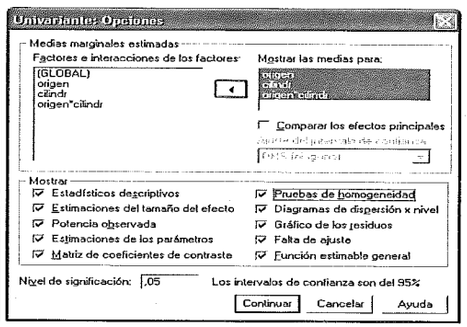


Figura 8-6

El botón *Suma de cuadrados* de la Figura 8-3 permite elegir entre cuatro métodos distintos para el cálculo de las sumas de cuadrados. Las *sumas de cuadrados Tipo I* (descomposición jerárquica) se caracterizan porque cada término se corrige sólo respecto al término que le precede en el modelo y se utiliza normalmente en modelos equilibrados en los que cualquier efecto principal se evalúa antes que cualquier efecto de interacción de primer orden y así sucesivamente. También se utiliza en los modelos anidados en los que el primer efecto especificado está anidado dentro del segundo efecto, el segundo se anida dentro del tercero y así sucesivamente.

Las *sumas de cuadrados Tipo II* se obtienen teniendo en cuenta únicamente los efectos que no están contenidos en el efecto que se está evaluando y se utilizan normalmente en los modelos equilibrados, en los modelos que sólo contienen efectos principales y en los diseños anidados en los que cada efecto especificado está anidado sobre el anterior. Las *sumas de cuadrados Tipo III* se obtienen ajustando cada efecto teniendo en cuenta cualquier otro efecto que no lo contenga y de forma independiente de cualquier efecto que lo contenga, si existe. Estas sumas de cuadrados, que son las más utilizadas, no se alteran por las variaciones del tamaño muestral de las casillas, de modo que son útiles especialmente para los modelos no equilibrados sin casillas vacías y para cualquier modelo para el que sean útiles las sumas de tipos I y II. Las *sumas de cuadrados Tipo IV* son útiles para analizar tanto modelos equilibrados como no equilibrados cuando existen casillas vacías.

Al pulsar *Aceptar* en la Figura 8-2 se obtiene la salida. La Figura 8-7 muestra el nombre de los factores, sus niveles con etiquetas de valores y el número de casos que hay en cado grupo. La Figura 8-8 presenta contrastes *ad hoc* que no aceptan la hipótesis nula de igualdad de varianzas para los distintos niveles de los factores. La Figura 8-9 muestra diversos estadísticos sobre los factores. La Figura 8-10 muestra la tabla resumen del ANCOVA, conteniendo fuentes de variación, sumas de cuadrados, grados de libertad, medias cuadráticas, estadísticos F y sus p-valores, referido todo ello a los efectos presentes en nuestro modelo de dos factores. También presenta medidas de la calidad del ajuste como la eta-cuadrado y otras medidas como el parámetro de no centralidad y la potencia de los contrastes. En general la significación de los parámetros del modelo resulta bastante alta (p-valores pequeños) salvo para *cilindr* y *origen* (con mayor p-valor y menor potencia).

Análisis de varianza univariante

Factores inter-sujetos

		Etiqueta del valor	N
País de origen	1	EE.UU.	249
	2	Europa	71
	3	Japón	79
Número de cilindros	3	3 cilindros	4
	4	4 cilindros	202
	5	5 cilindros	3
	6	6 cilindros	83
	8	8 cilindros	107

Figura 8-7

Estadísticos descriptivos

Variable dependiente: Potencia (CV)

País de origen	Número de cilindros	Media	Desv. típ.	N
EE.UU.	4 cilindros	80,96	10,540	69
	6 cilindros	99,67	12,935	73
	8 cilindros	158,13	27,871	107
	Total	119,61	39,799	249
Europa	4 cilindros	78,91	19,174	64
	5 cilindros	82,33	18,583	3
	6 cilindros	113,50	25,567	4
	Total	81,00	20,813	71
Japón	3 cilindros	99,25	8,302	4
	4 cilindros	75,58	13,983	69
	6 cilindros	115,83	12,106	6
	Total	79,84	17,819	79
Total	3 cilindros	99,25	8,302	4
	4 cilindros	78,47	14,967	202
	5 cilindros	82,33	18,583	3
	6 cilindros	101,51	14,310	83
	8 cilindros	158,13	27,871	107
	Total	104,86	38,566	390

Figura 8-8

Contraste de Levene sobre la igualdad de las varianzas error[a]

Variable dependiente: Potencia (CV)

F	gl1	gl2	Significación
2,309	8	390	,020

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intercept+peso+acel+motor+origen+cilindr+origen * cilindr

Figura 8-9

Pruebas de los efectos inter-sujetos

Variable dependiente: Potencia (CV)

Fuente		Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación	Eta al cuadrado parcial	Parámetro de no centralidad	Potencia observada(a)
Intersección	Hipótesis	18126,811	1	18126,811	127,121	,000	,343	127,121	1,000
	Error	34745,915	243,66	142,595(b)					
peso	Hipótesis	8218,506	1	8218,506	60,135	,000	,134	60,135	1,000
	Error	52890,089	387	136,667(c)					
acel	Hipótesis	24922,432	1	24922,432	182,359	,000	,320	182,359	1,000
	Error	52890,089	387	136,667(c)					
motor	Hipótesis	9517,310	1	9517,310	69,639	,000	,153	69,639	1,000
	Error	52890,089	387	136,667(c)					
origen	Hipótesis	5646,556	2	2823,278	2,090	,321	,673	4,181	,146
	Error	2743,686	2,031	1350,64(d)					
cilindr	Hipótesis	1231,091	4	307,773	,288	,867	,348	1,153	,065
	Error	2304,602	2,158	1067,86(e)					
origen * cilindr	Hipótesis	2902,892	2	1451,446	10,620	,000	,052	21,241	,989
	Error	52890,089	387	136,667(c)					

a Calculado con alfa = ,05
b ,040 MS(cilindr) - ,001 MS(origen * cilindr) + ,960 MS(Error)
c MS(Error)
d ,923 MS(origen * cilindr) + ,077 MS(Error)
e ,708 MS(origen * cilindr) + ,292 MS(Error)

Figura 8-10

La fila *Intersección* de la tabla ANCOVA (Figura 8-10) se refiere a la constante del modelo, necesaria para obtener las estimaciones de las medias de las casillas y contrasta la hipótesis de que la media total de la variable dependiente valga cero en la población, es decir la significatividad de la constante (como el p-valor es casi nulo, la significatividad es alta). Las filas siguientes recogen la significación de las covariables en el modelo, que también resulta alta (p-valores casi nulos). Las últimas filas recogen la significación de los efectos principales (efectos individuales de los factores del modelo) y del efecto de la interacción. Los factores no resultan significativos (p-valores altos), lo que indica que los grupos de coches definidos por las variables *origen* y *cilindr* poseen potencias no significativamente diferentes. El efecto de la interacción sí resulta significativo, lo que indica que las diferencias de potencia que se dan entre los automóviles de distinta cilindrada no son las mismas para los distintos orígenes considerados.

La Figura 8-11 muestra las estimaciones de los parámetros del modelo, a partir de las cuales se obtienen las medias que el modelo estima para cada nivel o combinaciones de niveles.

Las Figuras 8-12 a 8-18 presentan la matriz L de coeficientes del contraste que permite obtener los coeficientes asociados a cada efecto (coeficientes que definen el conjunto de hipótesis presentes en un determinado modelo).

Estimaciones de los parámetros

Variable dependiente: Potencia (CV)

Parámetro	B	Error típ.	T	Significación	Intervalo de confianza al 95%. Límite inferior	Límite superior	Eta cuadrado parcial	Parámetro de no centralidad	Potencia observada(a)
Intersección	86,233	8,863	9,729	,000	68,807	103,659	,197	9,729	1,000
peso	,051	,007	7,755	,000	,038	,064	,134	7,755	1,000
acel	-3,926	,291	-13,504	,000	-4,497	-3,354	,320	13,504	1,000
motor	,014	,002	8,345	,000	,011	,017	,153	8,345	1,000
[origen=1]	-26,440	5,350	-4,942	,000	-36,958	-15,92	,059	4,942	,999
[origen=2]	-,227	7,626	-,030	,976	-15,222	14,767	,000	,030	,050
[origen=3]	0(b)	.	.	.	.	.	.	.	.
[cilindr=3]	7,841	8,762	,895	,371	-9,386	25,068	,002	,895	,145
[cilindr=4]	-4,900	6,493	-,755	,451	-17,667	7,867	,001	,755	,117
[cilindr=5]	-16,256	9,559	-1,701	,090	-35,050	2,538	,007	1,701	,396
[cilindr=6]	-1,856	2,999	-,619	,536	-7,753	4,040	,001	,619	,095
[cilindr=8]	0(b)	.	.	.	.	.	.	.	.
[origen=1] * [cilindr=4]	21,211	5,617	3,777	,000	10,168	32,254	,036	3,777	,965
[origen=1] * [cilindr=6]	0(b)	.	.	.	.	.	.	.	.
[origen=1] * [cilindr=8]	0(b)	.	.	.	.	.	.	.	.
[origen=2] * [cilindr=4]	-,255	7,860	-,032	,974	-15,708	15,198	,000	,032	,050
[origen=2] * [cilindr=5]	0(b)	.	.	.	.	.	.	.	.
[origen=2] * [cilindr=6]	0(b)	.	.	.	.	.	.	.	.
[origen=3] * [cilindr=3]	0(b)	.	.	.	.	.	.	.	.
[origen=3] * [cilindr=4]	0(b)	.	.	.	.	.	.	.	.
[origen=3] * [cilindr=6]	0(b)	.	.	.	.	.	.	.	.

a Calculado con alfa = ,05

b Al parámetro se le ha asignado el valor cero porque es redundante.

Figura 8-11

Intersección

Parámetro	Contraste L1
Intersección	1,000
peso	,000
acel	,000
motor	,000
[origen=1]	,333
[origen=2]	,333
[origen=3]	,333
[cilindr=3]	,167
[cilindr=4]	,250
[cilindr=5]	,167
[cilindr=6]	,250
[cilindr=8]	,167
[origen=1] * [cilindr=4]	,083
[origen=1] * [cilindr=6]	,083
[origen=1] * [cilindr=8]	,167
[origen=2] * [cilindr=4]	,083
[origen=2] * [cilindr=5]	,167
[origen=2] * [cilindr=6]	,083
[origen=3] * [cilindr=3]	,167
[origen=3] * [cilindr=4]	,083
[origen=3] * [cilindr=6]	,083

Figura 8-12

peso

Parámetro	Contraste L2
Intersección	0
peso	1
acel	0
motor	0
[origen=1]	0
[origen=2]	0
[origen=3]	0
[cilindr=3]	0
[cilindr=4]	0
[cilindr=5]	0
[cilindr=6]	0
[cilindr=8]	0
[origen=1] * [cilindr=4]	0
[origen=1] * [cilindr=6]	0
[origen=1] * [cilindr=8]	0
[origen=2] * [cilindr=4]	0
[origen=2] * [cilindr=5]	0
[origen=2] * [cilindr=6]	0
[origen=3] * [cilindr=3]	0
[origen=3] * [cilindr=4]	0
[origen=3] * [cilindr=6]	0

Figura 8-13

acel

Parámetro	Contraste L3
Intersección	0
peso	0
acel	1
motor	0
[origen=1]	0
[origen=2]	0
[origen=3]	0
[cilindr=3]	0
[cilindr=4]	0
[cilindr=5]	0
[cilindr=6]	0
[cilindr=8]	0
[origen=1] * [cilindr=4]	0
[origen=1] * [cilindr=6]	0
[origen=1] * [cilindr=8]	0
[origen=2] * [cilindr=4]	0
[origen=2] * [cilindr=5]	0
[origen=2] * [cilindr=6]	0
[origen=3] * [cilindr=3]	0
[origen=3] * [cilindr=4]	0
[origen=3] * [cilindr=6]	0

Figura 8-14

motor

Parámetro	Contraste L4
Intersección	0
peso	0
acel	0
motor	1
[origen=1]	0
[origen=2]	0
[origen=3]	0
[cilindr=3]	0
[cilindr=4]	0
[cilindr=5]	0
[cilindr=6]	0
[cilindr=8]	0
[origen=1] * [cilindr=4]	0
[origen=1] * [cilindr=6]	0
[origen=1] * [cilindr=8]	0
[origen=2] * [cilindr=4]	0
[origen=2] * [cilindr=5]	0
[origen=2] * [cilindr=6]	0
[origen=3] * [cilindr=3]	0
[origen=3] * [cilindr=4]	0
[origen=3] * [cilindr=6]	0

Figura 8-15

origen

Parámetro	Contraste L5	Contraste L6
Intersección	,000	,000
peso	,000	,000
acel	,000	,000
motor	,000	,000
[origen=1]	1,000	,000
[origen=2]	,000	1,000
[origen=3]	-1,000	-1,000
[cilindr=3]	,000	,000
[cilindr=4]	,000	,000
[cilindr=5]	,000	,000
[cilindr=6]	,000	,000
[cilindr=8]	,000	,000
[origen=1] * [cilindr=4]	,500	,000
[origen=1] * [cilindr=6]	,500	,000
[origen=1] * [cilindr=8]	,000	,000
[origen=2] * [cilindr=4]	,000	,500
[origen=2] * [cilindr=5]	,000	,000
[origen=2] * [cilindr=6]	,000	,500
[origen=3] * [cilindr=3]	,000	,000
[origen=3] * [cilindr=4]	-,500	-,500
[origen=3] * [cilindr=6]	-,500	-,500

Figura 8-16

cilindr

Parámetro	Contraste L8	Contraste L9	Contraste L10	Contraste L11
Intersección	,000	,000	,000	,000
peso	,000	,000	,000	,000
acel	,000	,000	,000	,000
motor	,000	,000	,000	,000
[origen=1]	,000	,000	,000	,000
[origen=2]	,000	,000	,000	,000
[origen=3]	,000	,000	,000	,000
[cilindr=3]	1,000	,000	,000	,000
[cilindr=4]	,000	1,000	,000	,000
[cilindr=5]	,000	,000	1,000	,000
[cilindr=6]	,000	,000	,000	1,000
[cilindr=8]	-1,000	-1,000	-1,000	-1,000
[origen=1] * [cilindr=4]	,500	,667	,500	,333
[origen=1] * [cilindr=6]	,500	,333	,500	,667
[origen=1] * [cilindr=8]	-1,000	-1,000	-1,000	-1,000
[origen=2] * [cilindr=4]	,000	,167	-,500	-,167
[origen=2] * [cilindr=5]	,000	,000	1,000	,000
[origen=2] * [cilindr=6]	,000	-,167	-,500	,167
[origen=3] * [cilindr=3]	1,000	,000	,000	,000
[origen=3] * [cilindr=4]	-,500	,167	,000	-,167
[origen=3] * [cilindr=6]	-,500	-,167	,000	,167

Figura 8-17

origen * cilindr

Parámetro	Contraste L13	Contraste L16
Intersección	,000	,000
peso	,000	,000
acel	,000	,000
motor	,000	,000
[origen=1]	,000	,000
[origen=2]	,000	,000
[origen=3]	,000	,000
[cilindr=3]	,000	,000
[cilindr=4]	,000	,000
[cilindr=5]	,000	,000
[cilindr=6]	,000	,000
[cilindr=8]	,000	,000
[origen=1] * [cilindr=4]	1,000	,000
[origen=1] * [cilindr=6]	-1,000	,000
[origen=1] * [cilindr=8]	,000	,000
[origen=2] * [cilindr=4]	,000	1,000
[origen=2] * [cilindr=5]	,000	,000
[origen=2] * [cilindr=6]	,000	-1,000
[origen=3] * [cilindr=3]	,000	,000
[origen=3] * [cilindr=4]	-1,000	-1,000
[origen=3] * [cilindr=6]	1,000	1,000

Figura 8-18

Las Figuras 8-19 a 8-22 presentan los contrastes para la significatividad de las diferencias entre los niveles de cada factor con sus correspondientes resultados de la prueba (para el *contraste de la Diferencia* y el *contraste de Helmert* elegidos en la Figura 8-4). Se ve que la significación de las diferencias es bastante alta (p-valores pequeños). En el *contraste de la Diferencia* cada categoría, excepto la primera, se compara con la media de las categorías anteriores y en el *contraste de Helmert*, cada categoría, excepto la última, se compara con la media de las categorías posteriores. Las Figuras 8-23 a 8-28 muestran las medias marginales de la variable dependiente estimadas para cada nivel de factor y la matriz L^* de coeficientes del contraste que permite obtener los coeficientes asociados a cada efecto.

Resultados del contraste (matriz K)

Contraste de diferencias País de origen			Variable dependiente Potencia (CV)
Nivel 2 - Nivel 1	Estimación del contraste		15,480
	Valor hipotetizado		0
	Diferencia (Estimado - Hipotetizado)		15,480
	Error típ.		3,397
	Significación		,000
	Intervalo de confianza al 95 % para diferencia	Límite inferior	8,800
		Límite superior	22,159
Nivel 3 - Anterior	Estimación del contraste		8,095
	Valor hipotetizado		0
	Diferencia (Estimado - Hipotetizado)		8,095
	Error típ.		3,043
	Significación		,008
	Intervalo de confianza al 95 % para diferencia	Límite inferior	2,111
		Límite superior	14,078

Figura 8-19

Resultados de la prueba

Variable dependiente: Potencia (CV)

Fuente	Suma de cuadrados	gl	Media cuadrática	F	Significación	Eta al cuadrado parcial	Parámetro de no centralidad	Potencia observada[a]
Contraste	5646,556	2	2823,278	20,658	,000	,096	41,316	1,000
Error	52890,089	387	136,667					

a. Calculado con alfa = ,05

Figura 8-20

Resultados del contraste (matriz K)

Número de cilindros **contraste de Helmert**			Variable dependiente Potencia (CV)
Nivel 1 - Anterior	Estimación del contraste		8,291
	Valor hipotetizado		0
	Diferencia (Estimado - Hipotetizado)		8,291
	Error típ.		8,122
	Significación		,308
	Intervalo de confianza al 95 % para diferencia	Límite inferior	-7,677
		Límite superior	24,260
Nivel 2 - Anterior	Estimación del contraste		22,349
	Valor hipotetizado		0
	Diferencia (Estimado - Hipotetizado)		22,349
	Error típ.		4,447
	Significación		,000
	Intervalo de confianza al 95 % para diferencia	Límite inferior	13,606
		Límite superior	31,092
Nivel 3 - Anterior	Estimación del contraste		-15,328
	Valor hipotetizado		0
	Diferencia (Estimado - Hipotetizado)		-15,328
	Error típ.		9,146
	Significación		,095
	Intervalo de confianza al 95 % para diferencia	Límite inferior	-33,309
		Límite superior	2,653
Nivel 4 - Nivel 5	Estimación del contraste		-1,856
	Valor hipotetizado		0
	Diferencia (Estimado - Hipotetizado)		-1,856
	Error típ.		2,999
	Significación		,536
	Intervalo de confianza al 95 % para diferencia	Límite inferior	-7,753
		Límite superior	4,040

Figura 8-21

Resultados de la prueba

Variable dependiente: Potencia (CV)

Fuente	Suma de cuadrados	gl	Media cuadrática	F	Significación	Eta al cuadrado parcial	Parámetro de no centralidad	Potencia observada[a]
Contraste	6824,228	4	1706,057	12,483	,000	,114	49,933	1,000
Error	52890,089	387	136,667					

a. Calculado con alfa = ,05

Figura 8-22

Coeficientes de contraste (matriz L')

Parámetro	País de origen		
	EE.UU.	Europa	Japón
Intersección	1	1	1
peso	993,722	993,722	993,722
acel	15,485	15,485	15,485
motor	3203,566	3203,566	3203,566
[origen=1]	1	0	0
[origen=2]	0	1	0
[origen=3]	0	0	1
[cilindr=3]	0	0	,333
[cilindr=4]	,333	,333	,333
[cilindr=5]	0	,333	0
[cilindr=6]	,333	,333	,333
[cilindr=8]	,333	0	0
[origen=1] * [cilindr=4]	,333	0	0
[origen=1] * [cilindr=6]	,333	0	0
[origen=1] * [cilindr=8]	,333	0	0
[origen=2] * [cilindr=4]	0	,333	0
[origen=2] * [cilindr=5]	0	,333	0
[origen=2] * [cilindr=6]	0	,333	0
[origen=3] * [cilindr=3]	0	0	,333
[origen=3] * [cilindr=4]	0	0	,333
[origen=3] * [cilindr=6]	0	0	,333

Figura 8-23

Coeficientes de contraste (matriz L')

Parámetro	Número de cilindros				
	3 cilindros	4 cilindros	5 cilindros	6 cilindros	8 cilindros
Intersección	1	1	1	1	1
peso	993,722	993,722	993,722	993,722	993,722
acel	15,485	15,485	15,485	15,485	15,485
motor	3203,566	3203,566	3203,566	3203,566	3203,566
[origen=1]	0	,333	0	,333	1
[origen=2]	0	,333	1	,333	0
[origen=3]	1	,333	0	,333	0
[cilindr=3]	1	0	0	0	0
[cilindr=4]	0	1	0	0	0
[cilindr=5]	0	0	1	0	0
[cilindr=6]	0	0	0	1	0
[cilindr=8]	0	0	0	0	1
[origen=1] * [cilindr=4]	0	,333	0	0	0
[origen=1] * [cilindr=6]	0	0	0	,333	0
[origen=1] * [cilindr=8]	0	0	0	0	1
[origen=2] * [cilindr=4]	0	,333	0	0	0
[origen=2] * [cilindr=5]	0	0	1	0	0
[origen=2] * [cilindr=6]	0	0	0	,333	0
[origen=3] * [cilindr=3]	1	0	0	0	0
[origen=3] * [cilindr=4]	0	,333	0	0	0
[origen=3] * [cilindr=6]	0	0	0	,333	0

Figura 8-24

metro	EE.UU. Número de cilindros			Europa Número de cilindros			Japón Número de cilindros		
	4 cilindros	6 cilindros	8 cilindros	4 cilindros	5 cilindros	6 cilindros	3 cilindros	4 cilindros	6 cilindros
ección	1	1	1	1	1	1	1	1	1
	993,722	993,722	993,722	993,722	993,722	993,722	993,722	993,722	993,722
	15,485	15,485	15,485	15,485	15,485	15,485	15,485	15,485	15,485
r	3203,566	3203,566	3203,566	3203,566	3203,566	3203,566	3203,566	3203,566	3203,566
en=1]	1	1	1	0	0	0	0	0	0
en=2]	0	0	0	1	1	1	0	0	0
en=3]	0	0	0	0	0	0	1	1	1
dr=3]	0	0	0	0	0	0	1	0	0
dr=4]	1	0	0	1	0	0	0	1	0
dr=5]	0	0	0	0	1	0	0	0	0
dr=6]	0	1	0	0	0	1	0	0	1
dr=8]	0	0	1	0	0	0	0	0	0
en=1] * [cilindr=4]	1	0	0	0	0	0	0	0	0
en=1] * [cilindr=6]	0	1	0	0	0	0	0	0	0
en=1] * [cilindr=8]	0	0	1	0	0	0	0	0	0
en=2] * [cilindr=4]	0	0	0	1	0	0	0	0	0
en=2] * [cilindr=5]	0	0	0	0	1	0	0	0	0
en=2] * [cilindr=6]	0	0	0	0	0	1	0	0	0
en=3] * [cilindr=3]	0	0	0	0	0	0	1	0	0

Figura 8-25

Estimaciones

Variable dependiente: Potencia (CV)

País de origen	Media	Error típ.	Intervalo de confianza al 95%.	
			Límite inferior	Límite superior
EE.UU.	99,064[a,b]	1,061	96,979	101,149
Europa	112,703[a,b]	3,341	106,134	119,272
Japón	121,048[a,b]	3,214	114,729	127,366

a. Las covariables que aparecen en el modelo se evalúan en los siguiente valores: Peso total (kg) = 993,72, Aceleración 0 a 100 km/h (segundos) = 15,49, Cilindrada en cc = 3203,57.

b. Basada en la media marginal poblacional modificada.

Figura 8-26

Estimaciones

Variable dependiente: Potencia (CV)

Número de cilindros	Media	Error típ.	Intervalo de confianza al 95%.	
			Límite inferior	Límite superior
3 cilindros	128,527[a,b]	6,568	115,614	141,441
4 cilindros	113,882[a]	1,911	110,126	117,639
5 cilindros	104,203[a,b]	6,905	90,626	117,779
6 cilindros	109,941[a]	2,655	104,720	115,161
8 cilindros	94,246[a,b]	3,211	87,933	100,559

a. Las covariables que aparecen en el modelo se evalúan en los siguiente valores: Peso total (kg) = 993,72, Aceleración 0 a 100 km/h (segundos) = 15,49, Cilindrada en cc = 3203,57.

b. Basada en la media marginal poblacional modificada.

Figura 8-27

Estimaciones

Variable dependiente: Potencia (CV)

País de origen	Número de cilindros	Media	Error típ.	Intervalo de confianza al 95%. Límite inferior	Límite superior
EE.UU.	3 cilindros	.[a,b]	.	.	.
	4 cilindros	110,557[a]	2,009	106,606	114,508
	5 cilindros	.[a,b]	.	.	.
	6 cilindros	92,390[a]	1,570	89,302	95,477
	8 cilindros	94,246[a]	3,211	87,933	100,559
Europa	3 cilindros	.[a,b]	.	.	.
	4 cilindros	115,304[a]	2,331	110,721	119,887
	5 cilindros	104,203[a]	6,905	90,626	117,779
	6 cilindros	118,602[a]	6,059	106,689	130,515
	8 cilindros	.[a,b]	.	.	.
Japón	3 cilindros	128,527[a]	6,568	115,614	141,441
	4 cilindros	115,786[a]	2,394	111,080	120,493
	5 cilindros	.[a,b]	.	.	.
	6 cilindros	118,830[a]	4,936	109,125	128,535
	8 cilindros	.[a,b]	.	.	.

a. Las covariables que aparecen en el modelo se evalúan en los siguiente valores: Peso total (kg) = 993,72, Aceleración 0 a 100 km/h (segundos) = 15,49, Cilindrada en cc = 3203,57.

b. Esta combinación de niveles de los factores no tiene observaciones, por lo que la correspondiente media marginal poblacional no es estimable.

Figura 8-28

Las Figuras 8-29 y 8-30 presentan los *diagramas de dispersión por nivel* que proporcionan información gráfica sobre la igualdad de varianzas y que ayudan a detectar la posible existencia de algún tipo de relación entre el tamaño de las medias y el de las varianzas. Cuando las varianzas son iguales, los puntos del gráfico se muestran a la misma altura, es decir, alineados horizontalmente (hecho que no ocurre aquí y que coincide con el resultado previo del test de Levenne).

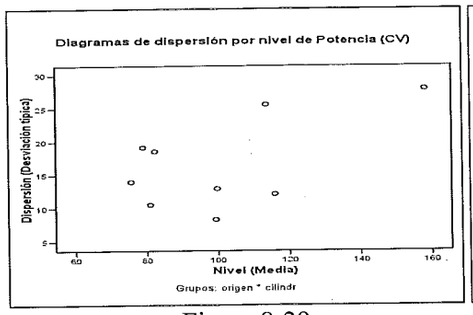


Figura 8-29

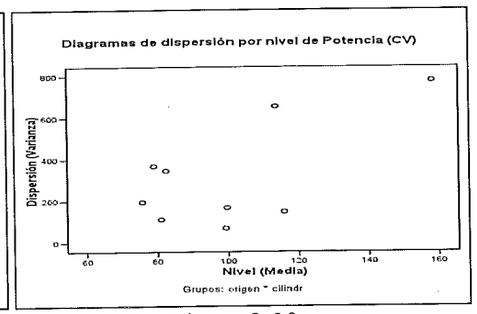


Figura 8-30

El *gráfico de los residuos* de la Figura 8-31 permite observar la aleatoriedad de los mismos y la independencia entre sí. Como el gráfico relativo a valores pronosticados y residuos tipificados es aleatorio, los residuos son independientes. Las varianzas residuales son homogéneas porque la dispersión de los residuos tipificados es similar a lo largo de todos los valores pronosticados. Valores pronosticados y observados muestran una pauta lineal lo que indica un buen ajuste lineal. El *gráfico de perfil de los efectos* (Figura 8-32) muestra que las líneas no se cortan claramente, luego la interacción no será demasiado significativa. Además, para todos ellos, la calidad del recuerdo decrece con el paso del tiempo hasta el segundo nivel, a partir del cual se ve una mejora.

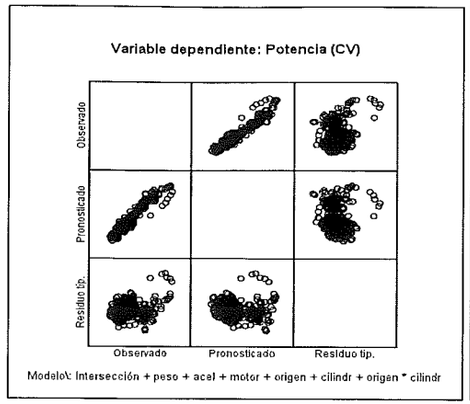


Figura 8-31

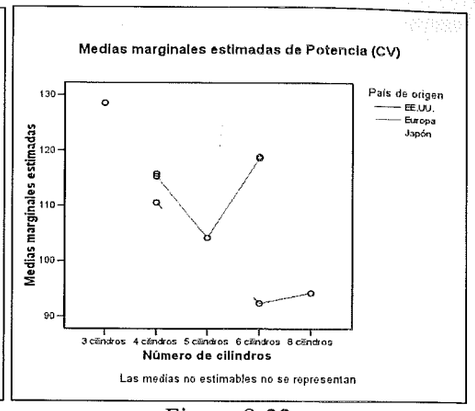


Figura 8-32

SPSS Y LA ESTIMACIÓN DE LAS COMPONENTES DE LA VARIANZA EN MODELOS ANCOVA DE EFECTOS MIXTOS

SPSS dispone del procedimiento *Componentes de la varianza*, que para modelos ANOVA y ANCOVA de efectos mixtos, estima la contribución de cada efecto aleatorio a la varianza de la variable dependiente. Este procedimiento resulta de particular interés para el análisis de modelos mixtos, como los diseños *split-plot*, los diseños de medidas repetidas univariados y los diseños de bloques aleatorios. Al calcular las componentes de la varianza, se puede determinar dónde centrar la atención para reducir la varianza. Se dispone de cuatro métodos diferentes para estimar las componentes de la varianza: estimador mínimo no cuadrático insesgado (EMNCI, MINQUE), análisis de varianza (ANOVA), máxima verosimilitud (MV, ML) y máxima verosimilitud restringida (MVR, RML). Se dispone de diversas especificaciones para los diferentes métodos. Los resultados por defecto para todos los métodos incluyen las estimaciones de componentes de la varianza. Si se usa el método MV o el método MVR, se mostrará también una tabla con la matriz de covarianza asintótica. Otros resultados disponibles incluyen una tabla de ANOVA o ANCOVA y las medias cuadráticas esperadas para el método ANOVA, y la historia de iteraciones para los métodos MV y MVR. El procedimiento *Componentes de la varianza* es totalmente compatible con el procedimiento *MLG Factorial general*. La opción *Ponderación MCP* permite especificar una variable usada para aplicar a las observaciones diferentes ponderaciones para un análisis ponderado; por ejemplo, para compensar la distinta precisión de las medidas.

Para realizar un análisis de componentes de la varianza, elija en los menús *Analizar* → *Modelo lineal general* → *Componentes de la varianza* (Figura 8-33), seleccione una variable dependiente y seleccione variables para *Factor(es) fijo(s), Factor(es) aleatorio(s)* y *Covariable(s)*, en función de los datos (Figura 8-34). Para especificar una variable de ponderación, utilice *Ponderación MCP*. Usaremos el mismo ejemplo del apartado anterior. Los botones *Modelo* (Figura 8-35) y *Método* (Figura 8-36) permiten elegir modelo y método de estimación. El botón *Guardar* permite salvar a archivo los resultados.

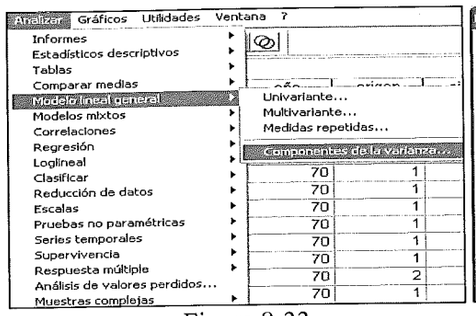


Figura 8-33

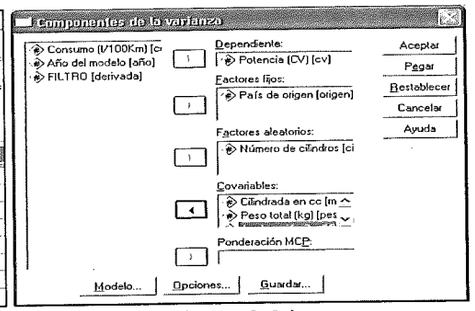


Figura 8-34

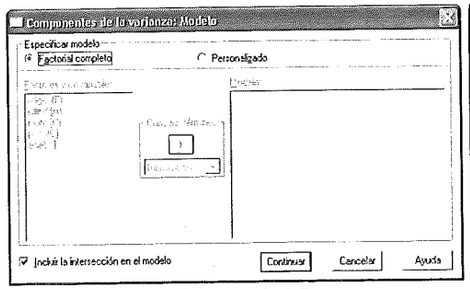


Figura 8-35

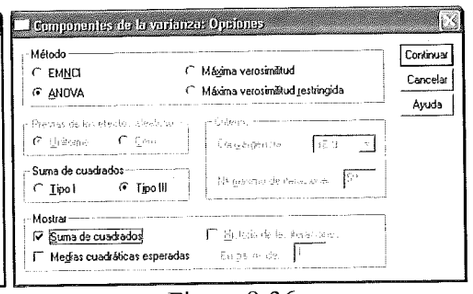


Figura 8-36

Al pulsar *Aceptar* en la Figura 8-34, se obtiene la salida, que presenta información sobre los niveles de los factores (Figura 8-37), la tabla ANOVA (Figura 8-38) y la estimación de las componentes de la varianza (Figura 8-39).

Información sobre los niveles de los factores

		Etiqueta de Valor	N
Número de cilindros	3	3 cilindros	4
	4	4 cilindros	202
	5	5 cilindros	3
	6	6 cilindros	83
	8	8 cilindros	107
País de origen	1	EE.UU.	249
	2	Europa	71
	3	Japón	79

Variable dependiente: cv

Figura 8-37

ANOVA

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática
Modelo corregido	539065,330	11	49005,939
Intersección	18126,811	1	18126,811
motor	9517,310	1	9517,310
peso	8218,506	1	8218,506
acel	24922,432	1	24922,432
cilindr	1231,091	4	307,773
origen	5646,556	2	2823,278
cilindr * origen	2902,892	2	1451,446
Error	52890,089	387	136,667
Total	4979388,000	399	
Total corregido	591955,419	398	

Variable dependiente: cv

Figura 8-38

Estimaciones de la varianza

Componente	Estimación
Var(cilindr)	-46,790[a]
Var(cilindr * origen)	93,094
Var(Error)	136,667

Variable dependiente: cv
Método: ANOVA (Tipo III Suma de cuadrados)
a. Con los métodos ANOVA y MINQUE pueden producirse estimaciones negativas de la componente de la varianza. Algunas razones para ello son: (a) el modelo especificado no es el correcto, o (b) el valor real de la varianza es igual a cero.

Figura 8-39

SPSS Y LOS MODELOS MANOVA Y MANCOVA MULTIVARIANTES DE UNO Y VARIOS FACTORES

SPSS incorpora el procedimiento *MLG Multivariante* que proporciona un análisis de regresión y un análisis de varianza y la covarianza para variables dependientes múltiples y para una o más covariables (cuantitativas) y variables de factor (cualitativas). Las calses de las variables de factor dividen la población en grupos. Utilizando este procedimiento del modelo lineal general, es posible contrastar hipótesis nulas sobre los efectos de las variables de factor sobre las medias de varias agrupaciones de una distribución conjunta de variables dependientes. Asimismo, se puede investigar las interacciones entre los factores y también los efectos individuales de los factores. Además, se pueden incluir los efectos de las covariables y las interacciones de covariables con los factores. Para el análisis de regresión, las variables independientes (predictoras) se especifican como covariables.

Se pueden contrastar tanto los modelos equilibrados como los no equilibrados. Se considera que un diseño está equilibrado si cada casilla del modelo contiene el mismo número de casos. En un modelo multivariado, las sumas de cuadrados debidas a los efectos del modelo y las sumas de cuadrados error se encuentran en forma de matriz en lugar de en la forma escalar del análisis univariado. Estas matrices se denominan matrices SCPC (sumas de cuadrados y productos cruzados). Si se especifica más de una variable dependiente, se proporciona el análisis multivariado de varianzas usando la traza de Pillai, la lambda de Wilks, la traza de Hotelling y el criterio de mayor raíz de Roy con el estadístico F aproximado, así como el análisis univariado de varianza para cada variable dependiente. Además de contrastar hipótesis, *MLG Multivariante* genera estimaciones de los parámetros. También se encuentran disponibles los contrastes *a priori* de uso más habitual para contrastar las hipótesis. Además, si una prueba F global ha mostrado cierta significación, pueden emplearse las pruebas *post hoc* para evaluar las diferencias entre las medias específicas. Las medias marginales estimadas ofrecen estimaciones de valores de las medias pronosticados para las casillas del modelo; los gráficos de perfil (gráficos de interacciones) de estas medias permiten observar fácilmente algunas de estas relaciones. Las pruebas de comparaciones múltiples *post hoc* se realizan por separado para cada variable dependiente.

Es posible guardar residuos, valores pronosticados, distancia de Cook y valores de influencia como variables nuevas para comprobar los supuestos. También se hallan disponibles una matriz SCPC residual, que es una matriz cuadrada de las sumas de cuadrados y los productos cruzados de los residuos; una matriz de covarianza residual, que es la matriz SCPC residual dividida por los grados de libertad de los residuos; y la matriz de correlaciones residual, que es la forma tipificada de la matriz de covarianza residual. Ponderación MCP permite especificar una variable usada para aplicar a las observaciones una ponderación diferencial en un análisis de mínimos cuadrados ponderados (MCP), por ejemplo para compensar la distinta precisión de las medidas.

En cuanto a estadísticos se obtienen las pruebas de rango *post hoc* y las comparaciones múltiples, diferencia menos significativa (DMS), Bonferroni, Sidak, Scheffé, múltiples F de Ryan-Einot-Gabriel-Welsch (R-E-G-W-F), rango múltiple de Ryan-Einot-Gabriel-Welsch, Student-Newman-Keuls (S-N-K), diferencia honestamente significativa de Tukey, b de Tukey, Duncan, *GT2* de Hochberg, Gabriel, pruebas *t* de Waller Duncan, Dunnett (unilateral y bilateral), *T2* de Tamhane, *T3* de Dunnett, Games-Howell y *C* de Dunnett, estadísticos descriptivos, medias observadas, desviaciones típicas y recuentos de todas las variables dependientes en todas las casillas; la prueba de Levene sobre la homogeneidad de la varianza; la prueba *M* de Box sobre la homogeneidad de las matrices de covarianza de la variables dependientes; y la prueba de esfericidad de Bartlett. En cuanto a gráficos se obtienen diagramas de dispersión por nivel, gráficos de residuos y gráficos de perfil (interacción).

Para realizar un análisis de varianza MLG Multivariante, elija en los menús *Analizar* → *Modelo lineal general* → *Multivariante* (Figura 8-40) y seleccione al menos dos variables dependientes. Si lo desea, puede especificar *Factores fijos, Covariables* y *Ponderación MCP* (Figura 8-41). Los botones *Modelos* (Figura 8-42), *Contrastes, Gráficos, Post hoc, Guardar* y *Opciones* (Figura 8-43) funcionan como en el MLG Univariante. Usaremos el ejemplo del procedimiento anterior añadiendo *consumo* como variable dependiente con los datos del archivo *coches.sav*.

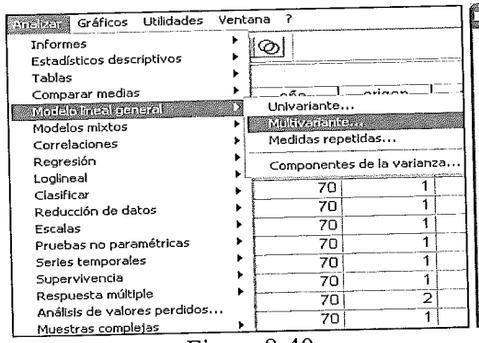


Figura 8-40

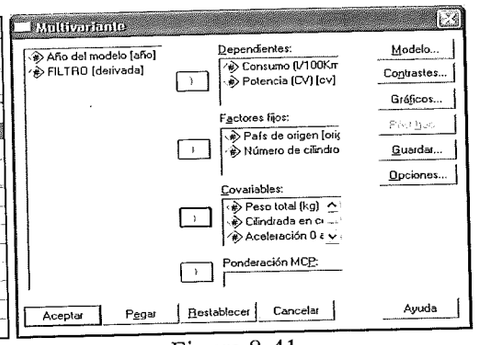


Figura 8-41

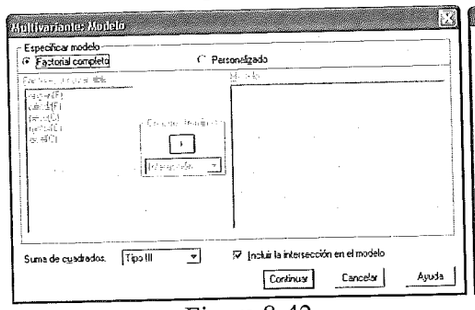


Figura 8-42

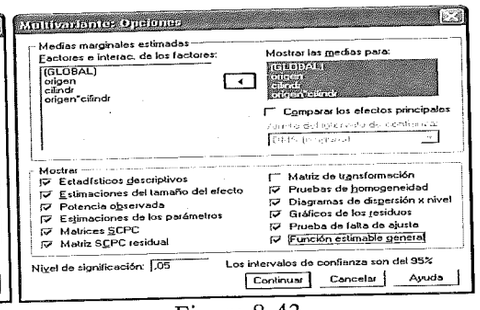


Figura 8-43

Al pulsar *Aceptar* en la Figura 8-41 se obtiene la salida. La Figura 8-44 muestra el nombre de los factores, sus niveles con etiquetas de valores y el número de casos que hay en cada grupo.

La Figura 8-45 presenta contrastes *ad hoc* para la igualdad de matrices de covarianzas y su proporcionalidad a la matriz identidad (*Prueba de esfericidad de Barlett*). La Figura 8-46 muestra diversos estadísticos sobre los factores.

La Figura 8-47 muestra diversos contrastes multivariados sobre la pertinencia del ajuste cuyos p-valores bajos en general admiten el ajuste. La Figura 8-48 muestra el estadístico de Levenne para la igualdad de varianzas del error (según su p-valor, se rechaza).

Las Figuras 8-49 a 8-52 muestran las matrices SCPC y las medias marginales estimadas de los factores y sus interaccionmes.

La Figura 8-53 muestra la tabla resumen del MANCOVA, conteniendo fuentes de variación, sumas de cuadrados, grados de libertad, medias cuadráticas, estadísticos *F* y sus p-valores, referido todo ello a los efectos presentes en nuestro modelo de dos factores. También presenta medidas de la calidad del ajuste como la eta-cuadrado y otras medidas como el parámetro de no centralidad y la potencia de los contrastes. En general la significación de los parámetros del modelo resulta bastante alta (p-valores pequeños) salvo para la interacción (con mayor p-valor y menor potencia).

La Figura 8-54 muestra las estimaciones de los parámetros del modelo, a partir de las cuales se obtienen las medias que el modelo estima para cada nivel o combinaciones de niveles.

Las Figuras 8-55 a 8-58 presentan los *diagramas de dispersión por nivel* que proporcionan información gráfica sobre la igualdad de varianzas, hipótesis que se rechaza porque los puntos del gráfico no se muestran a la misma altura (coincide con el resultado previo del test de Levenne).

Los *gráficos de los residuos* de las Figuras 8-59 y 8-60 permiten observar la aleatoriedad de los mismos y la independencia entre sí. Como los gráfico relativos a valores pronosticados y residuos tipificados son aleatorios, los residuos son independientes. Las varianzas residuales son homogéneas porque la dispersión de los residuos tipificados es similar a lo largo de todos los valores pronosticados. Valores pronosticados y observados muestran una pauta lineal lo que indica un buen ajuste lineal.

Modelo lineal general

Factores inter-sujetos

		Etiqueta del valor	N
País de origen	1	EE.UU.	244
	2	Europa	68
	3	Japón	79
Número de cilindros	3	3 cilindros	4
	4	4 cilindros	199
	5	5 cilindros	3
	6	6 cilindros	83
	8	8 cilindros	102

Figura 8-44

Prueba de Box sobre la igualdad de las matrices de covarianza[a]

M de Box	201,075
F	7,092
gl1	24
gl2	999,595
Significación	,000

Contrasta la hipótesis nula de que las matrices de covarianza observadas de las variables dependientes son iguales en todos los grupos.

a. Diseño: Intercept+peso+motor+acel+origen+cilindr+origen * cilindr

Prueba de esfericidad de Bartlett[a]

Razón de verosimilitud	,000
Chi-cuadrado aprox.	989,371
gl	2
Significación	,000

Contrasta la hipótesis nula de que la matriz de covarianza residual es proporcional a una matriz identidad.

a. Diseño: Intercept+peso+motor+acel+origen+cilindr+origen * cilindr

Figura 8-45

Estadísticos descriptivos

	País de origen	Número de cilindros	Media	Desv. típ.	N
Consumo (l/100Km)	EE.UU.	4 cilindros	8,64	1,361	69
		6 cilindros	12,26	1,795	73
		8 cilindros	16,22	2,738	102
		Total	12,89	3,806	244
	Europa	4 cilindros	8,70	1,820	61
		5 cilindros	9,00	3,000	3
		6 cilindros	12,75	3,202	4
		Total	8,96	2,147	68
	Japón	3 cilindros	11,50	1,291	4
		4 cilindros	7,67	1,492	69
		6 cilindros	10,17	1,941	6
		Total	8,05	1,825	79
	Total	3 cilindros	11,50	1,291	4
		4 cilindros	8,32	1,623	199
		5 cilindros	9,00	3,000	3
		6 cilindros	12,13	1,937	83
		8 cilindros	16,22	2,738	102
		Total	11,23	3,893	391
Potencia (CV)	EE.UU.	4 cilindros	80,96	10,540	69
		6 cilindros	99,67	12,935	73
		8 cilindros	157,96	28,383	102
		Total	118,75	39,696	244
	Europa	4 cilindros	78,31	18,219	61
		5 cilindros	82,33	18,583	3
		6 cilindros	113,50	25,567	4
		Total	80,56	20,158	68
	Japón	3 cilindros	99,25	8,302	4
		4 cilindros	75,58	13,983	69
		6 cilindros	115,83	12,106	6
		Total	79,84	17,819	79
	Total	3 cilindros	99,25	8,302	4
		4 cilindros	78,28	14,523	199
		5 cilindros	82,33	18,583	3
		6 cilindros	101,51	14,310	83
		8 cilindros	157,96	28,383	102
		Total	104,24	38,278	391

Figura 8-46

Contrastes multivariados(d)

Efecto		Valor	F	Gl hipótesis	Gl del error	Significación	Eta cuadrado parcial	Parámetro de no centralidad	Potencia observada(a)
Intersección	Traza de Pillai	,264	67,876(b)	2,000	378,000	,000	,264	135,753	1,000
	Lambda de Wilks	,736	67,876(b)	2,000	378,000	,000	,264	135,753	1,000
	Traza de Hotelling	,359	67,876(b)	2,000	378,000	,000	,264	135,753	1,000
	Raíz mayor de Roy	,359	67,876(b)	2,000	378,000	,000	,264	135,753	1,000
peso	Traza de Pillai	,187	43,601(b)	2,000	378,000	,000	,187	87,202	1,000
	Lambda de Wilks	,813	43,601(b)	2,000	378,000	,000	,187	87,202	1,000
	Traza de Hotelling	,231	43,601(b)	2,000	378,000	,000	,187	87,202	1,000
	Raíz mayor de Roy	,231	43,601(b)	2,000	378,000	,000	,187	87,202	1,000
motor	Traza de Pillai	,155	34,673(b)	2,000	378,000	,000	,155	69,346	1,000
	Lambda de Wilks	,845	34,673(b)	2,000	378,000	,000	,155	69,346	1,000
	Traza de Hotelling	,183	34,673(b)	2,000	378,000	,000	,155	69,346	1,000
	Raíz mayor de Roy	,183	34,673(b)	2,000	378,000	,000	,155	69,346	1,000
acel	Traza de Pillai	,328	92,107(b)	2,000	378,000	,000	,328	184,214	1,000
	Lambda de Wilks	,672	92,107(b)	2,000	378,000	,000	,328	184,214	1,000
	Traza de Hotelling	,487	92,107(b)	2,000	378,000	,000	,328	184,214	1,000
	Raíz mayor de Roy	,487	92,107(b)	2,000	378,000	,000	,328	184,214	1,000
origen	Traza de Pillai	,117	11,743	4,000	758,000	,000	,058	46,974	1,000
	Lambda de Wilks	,884	11,998(b)	4,000	756,000	,000	,060	47,992	1,000
	Traza de Hotelling	,130	12,251	4,000	754,000	,000	,061	49,005	1,000
	Raíz mayor de Roy	,122	23,069(c)	2,000	379,000	,000	,109	46,137	1,000
cilindr	Traza de Pillai	,070	3,422	8,000	758,000	,001	,035	27,378	,980
	Lambda de Wilks	,931	3,429(b)	8,000	756,000	,001	,035	27,428	,980
	Traza de Hotelling	,073	3,435	8,000	754,000	,001	,035	27,479	,980
	Raíz mayor de Roy	,055	5,193(c)	4,000	379,000	,000	,052	20,772	,968
origen * cilindr	Traza de Pillai	,068	6,620	4,000	758,000	,000	,034	26,481	,992
	Lambda de Wilks	,933	6,710(b)	4,000	756,000	,000	,034	26,839	,993
	Traza de Hotelling	,072	6,799	4,000	754,000	,000	,035	27,195	,994
	Raíz mayor de Roy	,070	13,324(c)	2,000	379,000	,000	,066	26,648	,998

a Calculado con alfa = ,05
b Estadístico exacto
c El estadístico es un límite superior para la F el cual ofrece un límite inferior para el nivel de significación.
d Diseño: Intercept+peso+motor+acel+origen+cilindr+origen * cilindr

Figura 8-47

Contraste de Levene sobre la igualdad de las varianzas error[a]

	F	gl1	gl2	Significación
Consumo (l/100Km)	4,242	8	382	,000
Potencia (CV)	2,491	8	382	,012

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intercept+peso+motor+acel+origen+cilindr+origen * cilindr

Figura 8-48

Matriz SCPC inter-sujetos

			Consumo (l/100Km)	Potencia (CV)
Hipótesis	Intersección	Consumo (l/100Km)	37,676	836,089
		Potencia (CV)	836,089	18554,158
	peso	Consumo (l/100Km)	175,926	1137,109
		Potencia (CV)	1137,109	7348,775
	motor	Consumo (l/100Km)	12,423	342,312
		Potencia (CV)	342,312	9432,101
	acel	Consumo (l/100Km)	9,750	486,938
		Potencia (CV)	486,938	24319,484
	origen	Consumo (l/100Km)	9,416	-25,491
		Potencia (CV)	-25,491	5488,849
	cilindr	Consumo (l/100Km)	59,714	114,050
		Potencia (CV)	114,050	1060,705
	origen * cilindr	Consumo (l/100Km)	3,269	-56,183
		Potencia (CV)	-56,183	2955,708
Error		Consumo (l/100Km)	1092,190	2374,083
		Potencia (CV)	2374,083	51665,736

Basado en la suma de cuadrados tipo III

Figura 8-49

Matriz SCPC residual

		Consumo (l/100Km)	Potencia (CV)
Suma de cuadrados y productos cruzados	Consumo (l/100Km)	1092,190	2374,083
	Potencia (CV)	2374,083	51665,736
Covarianza	Consumo (l/100Km)	2,882	6,264
	Potencia (CV)	6,264	136,321
Correlación	Consumo (l/100Km)	1,000	,316
	Potencia (CV)	,316	1,000

Basado en la suma de cuadrados tipo III

Figura 8-50

Matriz SCPC

		Consumo (l/100Km)	Potencia (CV)
Falta de ajuste	Consumo (l/100Km)	1087,690	2374,083
	Potencia (CV)	2374,083	51665,736
Error puro	Consumo (l/100Km)	4,500	,000
	Potencia (CV)	,000	,000

Medias marginales estimadas

1. Media global

Variable dependiente	Media	Error típ.	Intervalo de confianza al 95%. Límite inferior	Límite superior
Consumo (l/100Km)	11,275[a,b]	,230	10,823	11,727
Potencia (CV)	110,217[a,b]	1,580	107,109	113,324

a. Las covariables que aparecen en el modelo se evalúan en los siguiente valores: Peso total (kg) = 990,69, Cilindrada en cc = 3161,23, Aceleración 0 a 100 km/h (segundos) = 15,53.

b. Basada en la media marginal poblacional modificada.

Figura 8-51

2. País de origen

Variable dependiente	País de origen	Media	Error típ.	Intervalo de confianza al 95%. Límite inferior	Límite superior
Consumo (l/100Km)	EE.UU.	11,337[a,b]	,158	11,026	11,647
	Europa	10,848[a,b]	,484	9,896	11,801
	Japón	11,640[a,b]	,466	10,723	12,557
Potencia (CV)	EE.UU.	98,641[a,b]	1,085	96,507	100,774
	Europa	111,934[a,b]	3,332	105,382	118,486
	Japón	120,076[a,b]	3,207	113,771	126,381

a. Las covariables que aparecen en el modelo se evalúan en los siguiente valores: Peso total (kg) = 990,69, Cilindrada en cc = 3161,23, Aceleración 0 a 100 km/h (segundos) = 15,53.

b. Basada en la media marginal poblacional modificada.

3. Número de cilindros

Variable dependiente	Número de cilindros	Media	Error típ.	Intervalo de confianza al 95%. Límite inferior	Límite superior
Consumo (l/100Km)	3 cilindros	13,776[a,b]	,954	11,900	15,653
	4 cilindros	10,798[a]	,277	10,254	11,342
	5 cilindros	9,327[a,b]	1,002	7,356	11,298

Figura 8-52

Pruebas de los efectos inter-sujetos

Fuente	Variable dependiente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta cuadrado parcial	Parámetro de no centralidad	Potencia observada(a)
Modelo corregido	Consumo (l/100Km)	4818,552(b)	11	438,050	152,007	,000	,815	1672,082	1,000
	Potencia (CV)	519770,182(c)	11	47251,835	346,621	,000	,910	3812,834	1,000
Intersección	Consumo (l/100Km)	37,676	1	37,676	13,074	,000	,033	13,074	,950
	Potencia (CV)	18554,158	1	18554,158	136,106	,000	,264	136,106	1,000
peso	Consumo (l/100Km)	175,926	1	175,926	61,048	,000	,139	61,048	1,000
	Potencia (CV)	7349,775	1	7349,775	53,915	,000	,125	53,915	1,000
motor	Consumo (l/100Km)	12,423	1	12,423	4,311	,039	,011	4,311	,544
	Potencia (CV)	9432,101	1	9432,101	69,190	,000	,154	69,190	1,000
acel	Consumo (l/100Km)	9,750	1	9,750	3,383	,067	,009	3,383	,450
	Potencia (CV)	24319,484	1	24319,484	178,398	,000	,320	178,398	1,000
origen	Consumo (l/100Km)	9,416	2	4,708	1,634	,197	,009	3,267	,345
	Potencia (CV)	5488,849	2	2744,425	20,132	,000	,096	40,264	1,000
cilindr	Consumo (l/100Km)	59,714	4	14,929	5,180	,000	,052	20,721	,968
	Potencia (CV)	1060,705	4	265,176	1,945	,102	,020	7,781	,585
origen * cilindr	Consumo (l/100Km)	3,269	2	1,635	,567	,568	,003	1,134	,144
	Potencia (CV)	2955,708	2	1477,854	10,841	,000	,054	21,682	,990
Error	Consumo (l/100Km)	1092,190	379	2,882					
	Potencia (CV)	51665,736	379	136,321					
Total	Consumo (l/100Km)	55200,000	391						
	Potencia (CV)	4820275,000	391						
Total corregida	Consumo (l/100Km)	5910,742	390						
	Potencia (CV)	571435,918	390						

a Calculado con alfa = ,05
b R cuadrado = ,815 (R cuadrado corregida = ,810)
c R cuadrado = ,910 (R cuadrado corregida = ,907)

Figura 8-53

Estimaciones de los parámetros

Variable dependiente	Parámetro	B	Error típ.	t	Sig-nific.	Intervalo de confianza al 95%. Lím. Inf.	Límite sup	Eta cuadr. parcial	Parám. no centrali.	Poten. observada(a)
Consumo (l/100Km)	Intersección	3,152	1,303	2,420	,016	,591	5,713	,015	2,420	,675
	peso	,008	,001	7,813	,000	,006	,009	,139	7,813	1,000
	motor	,001	,000	2,076	,039	2,694E-05	,001	,011	2,076	,544
	acel	-,079	,043	-1,839	,067	-,163	,005	,009	1,839	,450
	[origen=1]	,894	,778	1,150	,251	-,635	2,424	,003	1,150	,209
	[origen=2]	1,528	1,108	1,379	,169	-,650	3,706	,005	1,379	,280
	[origen=3]	0(b)	.	.	.	.	.	.	.	.
	[cilindr=3]	2,772	1,274	2,175	,030	,266	5,278	,012	2,175	,583
	[cilindr=4]	-,413	,946	-,437	,662	-2,272	1,446	,001	,437	,072
	[cilindr=5]	-3,205	1,389	-2,307	,022	-5,936	-,474	,014	2,307	,634
	[cilindr=6]	-,452	,438	-1,032	,303	-1,313	,409	,003	1,032	,177
	[cilindr=8]	0(b)	.	.	.	.	.	.	.	.
	[origen=1] * [cilindr=4]	-,821	,816	-1,006	,315	-2,427	,784	,003	1,006	,171
	[origen=1] * [cilindr=6]	0(b)	.	.	.	.	.	.	.	.
	[origen=1] * [cilindr=8]	0(b)	.	.	.	.	.	.	.	.
	[origen=2] * [cilindr=4]	-,981	1,143	-,859	,391	-3,228	1,266	,002	,859	,137
	[origen=2] * [cilindr=5]	0(b)	.	.	.	.	.	.	.	.
	[origen=2] * [cilindr=6]	0(b)	.	.	.	.	.	.	.	.
	[origen=3] * [cilindr=3]	0(b)	.	.	.	.	.	.	.	.
	[origen=3] * [cilindr=4]	0(b)	.	.	.	.	.	.	.	.
	[origen=3] * [cilindr=6]	0(b)	.	.	.	.	.	.	.	.
Potencia (CV)	Intersección	89,01	8,959	9,936	,000	71,403	106,63	,207	9,936	1,000
	peso	,049	,007	7,343	,000	,036	,062	,125	7,343	1,000
	motor	,014	,002	8,318	,000	,011	,017	,154	8,318	1,000
	acel	-3,933	,294	-13,35	,000	-4,512	-3,354	,320	13,357	1,000
	[origen=1]	-26,21	5,351	-4,900	,000	-36,738	-15,69	,060	4,900	,998
	[origen=2]	,150	7,619	,020	,984	-14,832	15,131	,000	,020	,050
	[origen=3]	0(b)	.	.	.	.	.	.	.	.
	[cilindr=3]	6,825	8,765	,779	,437	-10,410	24,060	,002	,779	,121
	[cilindr=4]	-6,082	6,503	-,935	,350	-18,869	6,705	,002	,935	,154
	[cilindr=5]	-17,14	9,554	-1,795	,074	-35,932	1,639	,008	1,795	,433
	[cilindr=6]	-2,572	3,013	-,854	,394	-8,495	3,352	,002	,854	,136
	[cilindr=8]	0(b)	.	.	.	.	.	.	.	.
	[origen=1] * [cilindr=4]	21,16	5,615	3,770	,000	10,128	32,210	,036	3,770	,964
	[origen=1] * [cilindr=6]	0(b)	.	.	.	.	.	.	.	.
	[origen=1] * [cilindr=8]	0(b)	.	.	.	.	.	.	.	.
	[origen=2] * [cilindr=4]	-,904	7,860	-,115	,909	-16,358	14,551	,000	,115	,052
	[origen=2] * [cilindr=5]	0(b)	.	.	.	.	.	.	.	.
	[origen=2] * [cilindr=6]	0(b)	.	.	.	.	.	.	.	.
	[origen=3] * [cilindr=3]	0(b)	.	.	.	.	.	.	.	.
	[origen=3] * [cilindr=4]	0(b)	.	.	.	.	.	.	.	.
	[origen=3] * [cilindr=6]	0(b)	.	.	.	.	.	.	.	.

Figura 8-54

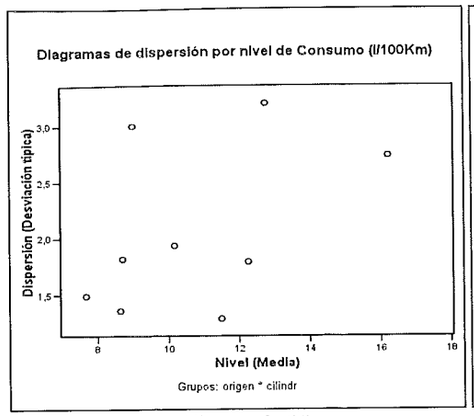


Figura 8-55

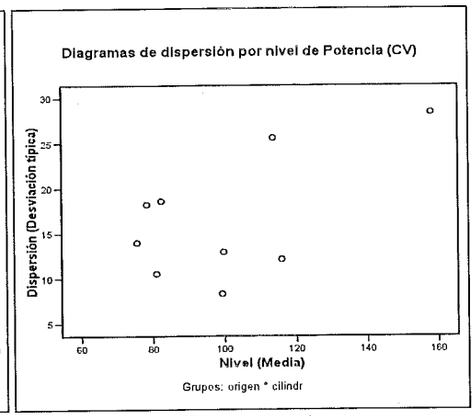


Figura 8-56

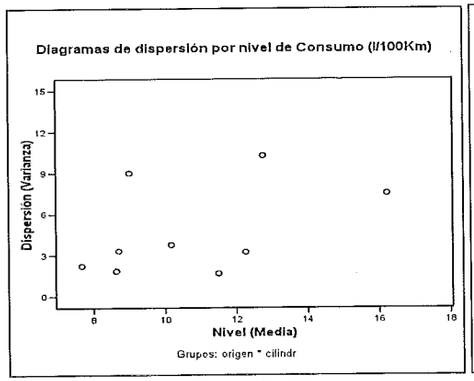


Figura 8-57

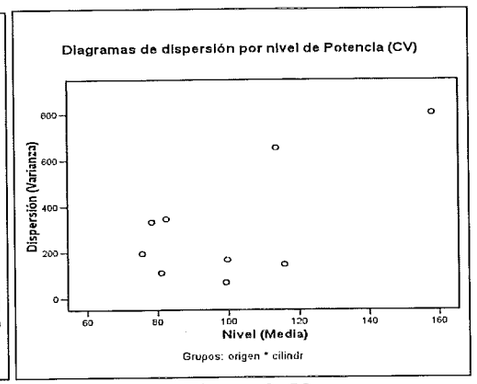


Figura 8-58

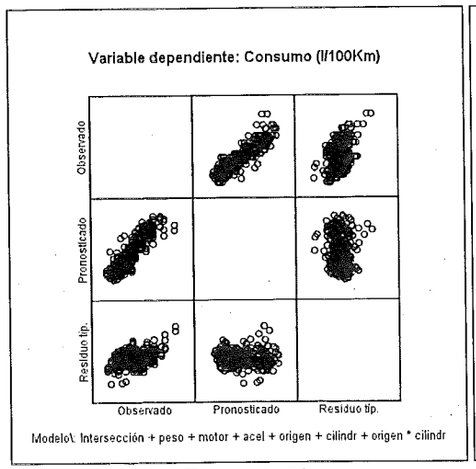


Figura 8-59

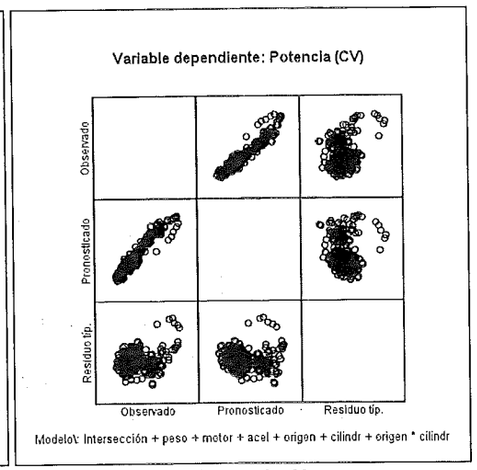


Figura 8-60

SPSS Y LOS MODELOS LINEALES MIXTOS

SPSS incorpora el procedimiento ***Modelos lineales mixtos*** que expande el modelo lineal general de modo que los datos puedan presentar variabilidad correlacionada y no constante. El modelo lineal mixto proporciona, por tanto, la flexibilidad necesaria para modelar no sólo las medias sino también las varianzas y covarianzas de los datos. Se trata por tanto, de un modelo utilizable en el caso de modelo de datos de panel.

La variable dependiente debe ser cuantitativa. Los factores deben ser categóricos y pueden tener valores numéricos o valores de cadena. Las covariables y la variable de ponderación deben ser cuantitativas. Las variables de sujetos y repetidas pueden ser de cualquier tipo. Se asume que la variable dependiente está relacionada linealmente con los factores fijos, los factores aleatorios y las covariables. Los efectos fijos modelan la media de la variable dependiente. Los efectos aleatorios modelan la estructura de las covarianzas de la variable dependiente. Los efectos aleatorios múltiples se consideran independientes entre sí y se calculan por separado las matrices de covarianzas de cada uno de ellos; sin embargo, se puede establecer una correlación entre los términos del modelo especificados para el mismo efecto aleatorio. Las medidas repetidas modelan la estructura de las covarianzas de los residuos. Se asume además que la variable dependiente procede de una distribución normal. Si no está claro que haya una variabilidad correlacionada o no constante, puede usarse el procedimiento MLG Univariante o MLG Medidas repetidas. Alternativamente, puede usarse el procedimiento *Análisis de componentes de la varianza* en caso de que los efectos aleatorios tengan una estructura de covarianzas en los componentes de la varianza y no haya medidas repetidas.

Como ejemplo, a partir del fichero *autos.sav*, ajustamos un modelo de panel de efectos fijos según el factor marca (*manufact*) que haga depender el precio (*price*) de los automóviles de su consumo (*mpg*) y de su potencia (*horsepower*). Para ejecutar el procedimiento, elija en los menús *Analizar* → *Modelos mixtos* → *Lineal...* (Figura 8-61). Si lo desea, seleccione variables de sujetos y repetidas en la Figura 8-62 (no es nuestro caso) y, a continuación, pulse en *Continuar*. En la Figura 8-63 seleccione una variable dependiente (*price*) y seleccione al menos un factor o covariable. A continuación pulse en *Fijos* (Figura 8-64) o *Aleatorios* y especifique al menos un modelo de efectos fijos (en nuestro caso sólo los efectos principales o variables independientes del modelo) o aleatorios. Si lo desea, seleccione una variable de ponderación (no es nuestro caso). El botón *Estimación* permite fijar determinadas características para la estimación del modelo (Figura 8-65). El botón *Estadísticos* permite elegir determinados estadísticos a obtener en la salida (Figura 8-66). Al pulsar *Continuar* y *Aceptar* se obtiene el panel de efectos fijos ajustado.

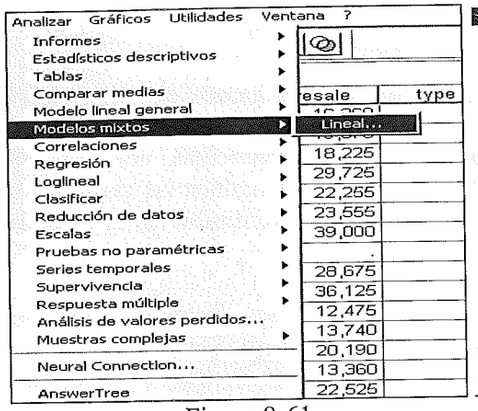


Figura 8-61

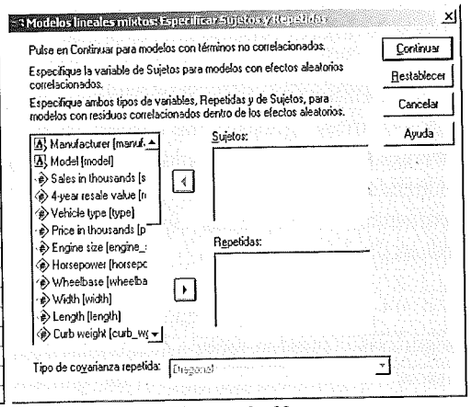


Figura 8-62

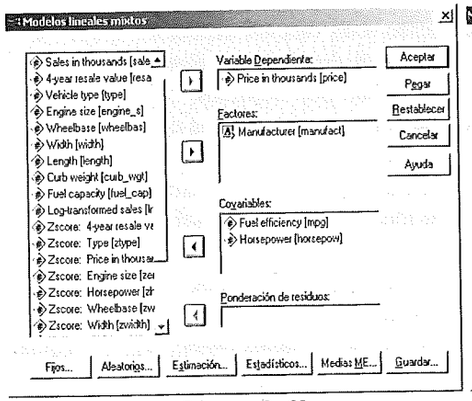


Figura 8-63

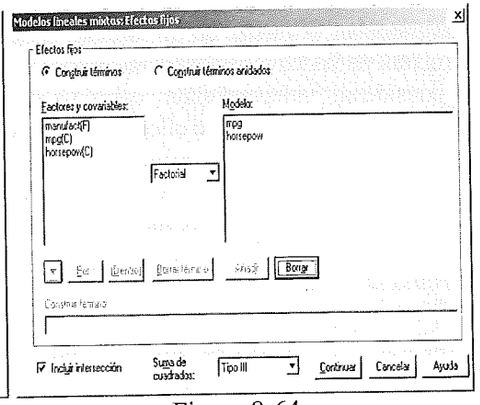


Figura 8-64

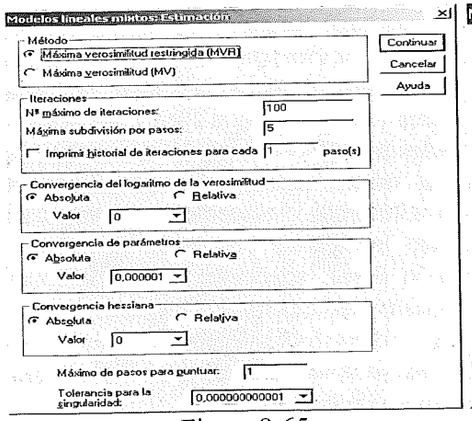


Figura 8-65

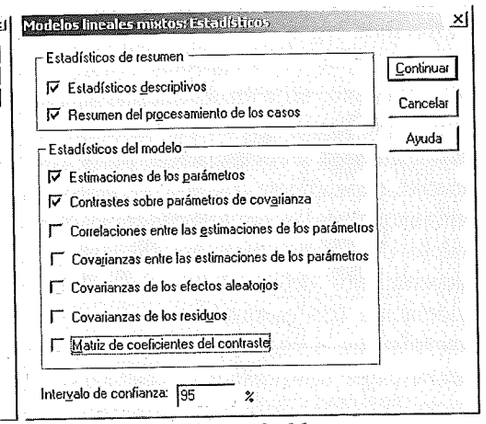


Figura 8-66

Análisis de modelos mixtos

Dimensión de modelo[a]

		Número de niveles	Número de parámetros
Efectos fijos	Intersección	1	1
	mpg	1	1
	horsepow	1	1
Residuos			1
Total		3	4

a. Variable dependiente: Price in thousands.

Criterios de información[a]

-2 log de la verosimilitud restringida	1068,017
Criterio de información de Akaike (AIC)	1070,017
Criterio de Hurvich y Tsai (AICC)	1070,045
Criterio de Bozdogan (CAIC)	1074,028
Criterio bayesiano de Schwarz (BIC)	1073,028

Figura 8-67

Tipo III de contrastes de efectos fijos[a]

Origen	Numerador df	Denominador df	Valor F	Sig.
Intersección	1	150	6,649	,011
mpg	1	150	,562	,455
horsepow	1	150	243,222	,000

a. Variable dependiente: Price in thousands.

Estimaciones de efectos fijos[a]

Parámetro	Estimación	Error típico	gl	t	Sig.	Intervalo de confianza 95% Límite inferior	Límite superior
Intersección	-16,599059	6,4373076	150	-2,579	,011	-29,3185685	-3,87955
mpg	,1403016	,1872245	150	,749	,455	-,2286363	,5102396
horsepow	,2195249	,0140761	150	15,596	,000	,1917118	,2473379

a. Variable dependiente: Price in thousands.

Estimaciones de parámetros de covarianza[a]

Parámetro	Estimación	Error típico	Wald Z	Sig.	Intervalo de confianza 95% Límite inferior	Límite superior
Residuos	61,0400241	7,0492949	8,660	,000	48,6772918	76,54256

Figura 8-68

Estimaciones de efectos fijos(b)

Parámetro	Estimación	Error típico	gl	t	Sig.	Intervalo de confianza 95% Límite inferior	Límite superior
Intersección	6,1865839	6,4326389	121,000	,962	,338	-6,5485216	18,9216894
mpg	-,2272574	,1704173	121	-1,334	,185	-,5646434	,1101286
horsepow	,1664111	,0133820	121,000	12,435	,000	,1399179	,1929043
[manufact=Acura]	-1,7672768	4,0480504	121	-,437	,663	-9,7814602	6,2469067
[manufact=Audi]	2,4856465	4,0646809	121	,612	,542	-5,5614614	10,5327544
[manufact=BMW]	1,7424183	4,0450223	121	,431	,667	-6,2657704	9,7506070
[manufact=Buick]	-8,2166316	3,7031996	121	-2,219	,028	-15,5480920	-,8851713
[manufact=Cadillac]	-3,8523232	3,5493600	121	-1,085	,280	-10,8792176	3,1745711
[manufact=Chevrolet]	-8,1743843	3,0715430	121	-2,661	,009	-14,2553138	-2,0934548
[manufact=Chrysler]	-9,6103739	3,3058561	121	-2,907	,004	-16,1551880	-3,0655598
[manufact=Dodge]	-10,3999426	3,0203871	121	-3,443	,001	-16,3795955	-4,4202897
[manufact=Ford]	-8,2587261	2,9305419	121	-2,818	,006	-14,0605068	-2,4569455
[manufact=Honda]	-6,9204897	3,4715185	121	-1,994	,048	-13,7932763	-,0477032
[manufact=Hyundai]	-8,8122023	4,0863694	121	-2,156	,033	-16,9022484	-,7221563
[manufact=Infiniti]	-8,8154689	6,2112881	121	-1,419	,158	-21,1123520	3,4814142
[manufact=Jaguar]	1,4471573	6,2076732	121	,233	,816	-10,8425691	13,7368837
[manufact=Jeep]	-8,9653145	4,1951651	121	-2,137	,035	-17,2707503	-,6598786
[manufact=Lexus]	1,6935497	3,3704097	121	,502	,616	-4,9790654	8,3661647
[manufact=Lincoln]	-3,9245306	4,1382835	121	-,948	,345	-12,1173543	4,2682931
[manufact=Mercedes-Benz]	13,4136505	3,0685055	121	4,371	,000	7,3387345	19,4885666
[manufact=Mercury]	-7,3235110	3,3201331	121	-2,206	,029	-13,8965902	-,7504318
[manufact=Mitsubishi]	-6,3063053	3,2140837	121	-1,962	,052	-12,6694318	,0568212
[manufact=Nissan]	-6,8847523	3,2213089	121	-2,137	,035	-13,2621829	-,5073218
[manufact=Oldsmobile]	-5,8283175	3,4705584	121	-1,679	,096	-12,6992032	1,0425682
[manufact=Plymouth]	-5,5548859	3,6947687	121	-1,503	,135	-12,8696551	1,7598834
[manufact=Pontiac]	-8,3799123	3,3026161	121	-2,537	,012	-14,9183121	-1,8415125
[manufact=Porsche]	15,9671225	4,1657299	121	3,833	,000	7,7199614	24,2142835
[manufact=Saab]	-,8876343	4,6785509	121	-,190	,850	-10,1500599	8,3747913
[manufact=Saturn]	-4,9097271	3,5585950	121	-1,380	,170	-11,9549044	2,1354503
[manufact=Subaru]	-6,6816095	4,6746848	121	-1,429	,155	-15,9363811	2,5731622
[manufact=Toyota]	-5,2084755	3,0228372	121	-1,723	,087	-11,1929791	,7760281
[manufact=Volkswagen]	-2,6496903	3,3714609	121	-,786	,433	-9,3243865	4,0250059
[manufact=Volvo]	0(a)	0	.	.	.	.	.

a Se ha establecido este parámetro en cero porque es redundante.
b Variable dependiente: Price in thousands.

Figura 8-69

La Figura 8-67 muestra varios criterios de información para valorar la calidad del ajuste. La Figura 8-68 muestra los parámetros estimados y su significatividad. Se observa que la variable consumo resulta no significativa, lo que puede llevarnos a intentar estimar el modelo suponiendo efectos aleatorios. Para obtener las estimaciones de los efectos fijos (Figura 8-69) se introduce la variable *Manufact* en el campo *Modelo* en la Figura 8-64 y se ejecuta el procedimiento.

ANÁLISIS DE LA VARIANZA Y LA COVARIANZA CON SAS: PROCEDIMIENTO GLM

El procedimiento GLM permite ejecutar regresión simple, regresión múltiple, análisis de la varianza y la covarianza simple y múltiple, modelos de superficie de respuesta, análisis multivariante de la varianza, análisis de la varianza con medidas repetidas, etc. PROC GLM permite trabajar con diseños no balanceados (ANOVA sólo permitía diseños balanceados). Su sintaxis resumida es la siguiente:

PROC GLM opciones;
CLASS variables;
MODEL dependientes=independientes / opciones;
ABSORB variables;
BY variables;
FREQ variable;
ID variables;
WEIGHT variable;
CONTRAST 'etiqueta' efecto values ... efecto values / opciones;
ESTIMATE 'etiqueta' efecto values ... efecto values / opciones;
LSMEANS efectos / opciones;
MANOVA test-opciones / detalle-opciones;
MEANS efectos / opciones;
OUTPUT OUT=conjunto de datos de salida estadísticos = nombres
RANDOM efectos / opciones;
REPEATED factor / opciones;
TEST H=efectos E=efecto / opciones;

La sentencia ABSORB absorbe efectos de clasificación en el modelo, BY especifica variable para la definición de subgrupos, CLASS declara variables de clasificación, CONTRAST contrasta funciones lineales de los parámetros, ESTIMATE estima funciones lineales de los parámetros, FREQ especifica una variable con frecuencias absolutas para las observaciones, ID identifica observaciones en la salida, LSMEANS halla medias por mínimo cuadrados, MANOVA ejecuta análisis multivariante de la varianza, MEANS compara medias aritméticas, MODEL define el modelo a ajustar, OUTPUT define el conjunto de datos de salida donde se guardan los resultados de los contrastes, RANDOM declara ciertos efectos como aleatorios y halla cuadrados medios esperados, REPEATED ejecuta análisis de la varianza univariante y multivariante en medidas repetidas, TEST construye tests y WEIGHT especifica una posible variable con los pesos de las observaciones.

Las **opciones de PROC GLM** son: DATA=conjunto de datos de entrada, ALPHA=nivel de significación, MANOVA para análisis multivariante de la varianza, MULTIPASS para que GLM relea datos de entrada las veces que sea necesario, NAMELEN=longitud de los nombres de los efectos en las tablas, NOPRINT que suprime la salida normal de resultados, ORDER=DATA | FREQ FORMATTED | INTERNAL y OUTSTAT=conjunto de datos de salida que contiene sumas de cuadrados, grados de libertad, estadístico F, niveles de probabilidad para cada efecto, MSE y otros resultados del ajuste.

Las **opciones de la sentencia MODEL** son las siguientes:

ALIASING (contrastes para la constante)	ALPHA=nivel de significación	CLI (límites de confianza para predicciones individuales de cada observación)
CLM (límites de confianza para la media y valores predichos)	CLPARM (límites de confianza para los parámetros estimados)	E (forma general para todas las funciones estimables)
E1 (funciones estimables tipo I para cada efecto y sumas de cuadrados)	E2(funciones estimables tipo II para cada efecto y sumas de cuadrados)	E3 ((funciones estimables tipo III para cada efecto y sumas de cuadrados)
E4 (funciones estimables tipo IV para cada efecto y sumas de cuadrados)	INTERCEPT (tests de hipótesis con constante en el modelo)	INVERSE (muestra la inversa generalizada de X'X)
NOINT (modelo sin constante)	NOUNI (no se muestran estadísticos univariantes)	P (muestra valores observados, predichos y residuos para cada observación)
SINGULAR=number (ajusta la singularidad)	SOLUTION (solución de valores normales para los parámetros estimados)	SS1 (sumas de cuadrados asociadas con funciones estimables tipo I para cada efecto)
SS2 (sumas de cuadrados asociadas con funciones estimables tipo II para cada efecto)	SS3 (sumas de cuadrados asociadas con funciones estimables tipo III para cada efecto)	SS4 (sumas de cuadrados asociadas con funciones estimables tipo IV para cada efecto)
TOLERANCE (tolerancia en la rutina SWEEP)	XPX (matriz de productos cruzados aumentada de X'X)	ZETA=value (ajusta la sensibilidad en funciones tipo III y IV

Las **opciones de la sentencia CONTRAST** son E (vector L completo), E=efecto del modelo como término de error, ETYPE=1,2,3,4 (contrastes de tipo I, II, III o IV), SINGULAR= nivel de ajuste de la singularidad.

Las **opciones de la sentencia ESTIMATE** son DIVISOR (valor por el que se dividen todos los coeficientes), E y SINGULAR.

Las **opciones de la sentencia LSMEANS** son las siguientes:

ADJUST= BON | DUNNET SCHEFFE, SIDAK | GT2 |SIMULATE | TUKEY | T para definir el método de ajuste	ALPHA=nivel de significación para los intervalos de confianza	AT variable = valor permite modificar los valores de las covariables
BYLEVEL define niveles BY	CL define límites de confianza para medias individuales	COV incluye varianzas y covarianzas en la salida
E muestra los coeficientes de las funciones lineales usadas	E=efecto del modelo a usar como término de error	ETYPE=1,2,3,4 correspondiendo con tests de tipos I, II, III y IV
NOPRINT suprime la salida normal de resultados	OBSMARGINS especifica ponderación potencial diferente para hallar los coeficientes	OUT=conjunto de datos de salida con valores, errores y covarianzas
PDIFF asigna p-valores para las diferencias de media	SLICE = efecto fijo	SINGULAR=nivel de singularidad
STDERR produce errores estándar	TDIFF produce valores de la T para todas las hipótesis	

Las **opciones de la sentencia MEANS** son las siguientes:

ALPHA=nivel de significación	BON ejecuta el test de Bonferroni de diferencia de medias
CLDIFF presenta intervalos de confianza para todos los pares de diferencias de medias	CLM presenta intervalos de confianza para la media de cada nivel
DEPONLY muestra sólo medias para las variables dependientes	DUNCAN ejecuta el contraste de rangos múltiples de Duncan
DUNNETT ejecuta el test de Dunnet de dos lados para la significatividad de tratamientos	DUNNETTL ejecuta el test de Dunnet de un lado para ver si su tratamiento es menor o igual que un valor de control
DUNNETTU ejecuta el test de Dunnet de un lado para ver si un tratamiento es mayor o igual que un valor de control	E=efecto especifica el error cuadrático medio utilizado en comparaciones múltiples
ETYPE=1,2,3,4 especifica el tipo de cuadrados medios para el efecto del error	GABRIEL ejecuta el procedimiento de comparaciones múltiples de Gabriel
GT2 ejecuta comparaciones pareadas	HOVTEST =BARLET | BF | LEVENE | OBRIEN ejecuta los respectivos contrastes de homogeneidad de varianzas
HTYPE=n (tipo MS para hipótesis MS)	KRATIO=valor del ratio type1/type2 del contraste de Duncan
LINES lista medias en orden descendente e indica subconjuntos no significativos	LSD ejecuta tests pareados de la T
NOSORT ausencia de orden	REGWQ ejecuta test de rangos múltiples de Ryan-Einot-Gabriel-Welsch en todas las medias de efectos principales
SCHEFFE ejecuta el procedimiento de Scheffe en todas las medias de efectos principales	SIDAK ejecuta los test pareados de diferencias de medias de Sidak

SMM ejecuta los test pareados de diferencias de medias basados en el módulo máximo estudentizado de Sidak	SNK ejecuta el test de rangos múltiples de Student-Newman-Keuls
T ejecuta tests pareados de mímina diferencia significativa de Fisher (LSD)	TUKEY ejecuta el test de rango estudentizado de Tukey en todas las medias de efectos principales
WALLER ejecuta el ratio de Waller-Duncan en todas las medias de efectos principales	WELCH ejecuta ANOVA simple de varianza ponderada de Welch

Las **opciones de la sentencia MANOVA** son: H=efectos | INTERCEPT | _ALL_ especifica efectos en el modelo, E=efecto especifica efecto del error, M=ecuación,...,ecuación | (fila de matriz,...,fila de matriz) especifica una matriz de transformación para las variables dependientes listadas en el modelo, MNAMES=nombres especifica nombres para las variables definidas por las ecuaciones en la especificación M=, PREFIX=nombre también especifica nombres para las variables transformadas definidas en la especificación M= , CANONICAL realiza análisis canónico de las matrices en vez de mostrar únicamente raíces características y vectores propios, ETYPE=1,2,3,4 especifica contrastes tipo I, II, III y IV para la matriz E, HTYPE=1,2,3,4 especifica contrastes tipo I, II, III y IV para la matriz H, ORTH requiere la ortonormalización por filas de la matriz de transformación indicada en M= antes de realizar el análisis, PRINTE muestra la matriz E de error SSCP, PRINTH muestra la matriz H de la hipótesis SSCP asociada con cada efecto especificado en H=, SUMMARY produce tablas del análisis de la varianza para cada variable independiente.

Las **opciones de la sentencia OUTPUT** son ALPHA=nivel de significación y OUT=conjunto de datos de salida y los **estadísticos a utilizar en la sentencia OUTPUT** son los siguientes:

COOKD: D de Cook para la influencia	COVRATIO: Influencia estándar de observación en la covarianza de los parámetros estimados	DFFITS: Influencia estándar de observación en valor predicho
H: Medida Leverage para la influencia	LCL: Límite inferior del intervalo de confianza para predicción individual	LCLM: Límite inferior del intervalo de confianza para predicción en media
PREDICTED | P: Valores predichos	PRESS: Residuo para una observación resultante de eliminarla del modelo	RESIDUAL | R: Residuo calculado como ACTUAL-PREDICHO
RSTUDENT: Residuo estudentizado con la observación corriente borrada	STDI: Error estándar de predicciones individuales	STDP. Error estándar de predicciones en media
STDR: Error estándar del residuo	STUDENT: Residuos estudentizados	UCL: Límite superior del intervalo de confianza para predicción individual

UCLM: Límite superior del intervalo de confianza para predicción en media		

Las opciones de la **sentencia RANDOM** son Q (muestra todas las formas cuadráticas en los efectos fijos que aparecen en los cuadrados medios esperados) y TEST (ejecuta test de hipótesis para cada efecto especificado en el modelo).

Las **opciones de la sentencia REPETEAD** son las siguientes:

CONTRAST (*ordinal-referencia-nivel*) genera contrastes entre niveles de factor y niveles de referencia	**HELMERT** genera contrastes entre cada nivel de factor y la media de los niveles subsecuentes
IDENTITY genera la transformación identidad correspondiente al factor asociado	**MEAN** (*ordinal-referencia-nivel*) genera contrastes entre niveles de factor y la media de todos los otros niveles de factor
POLYNOMIAL genera contrastes polinomiales ortogonales	**PROFILE** genera contrastes entre niveles adyacentes y el factor

Las opciones de la sentencia TEST son: ETYPE=1,2,3,4 especifica el tipo de sumas de cuadrados a utilizar para el término del error (I, II, III o IV) y HTYPE=1,2,3,4 especifica el tipo de sumas de cuadrados a utilizar para las hipótesis (I, II, III o IV).

Como primer ejemplo consideremos el contenido en hierro de determinadas aleaciones y su pérdida de peso al someterlas a un proceso de análisis de la corrosión. Se trata de comprobar si es razonable sostener que la pérdida de peso es una función cuadrática del contenido en hierro.

```
title 'Regression con PROC GLM';
   data aleacion;
      input hierro peso @@;
      datalines;
   0.01 127.6   0.48 124.0   0.71 110.8   0.95 103.9
   1.19 101.5   0.01 130.1   0.48 122.0   1.44  92.3
   0.71 113.1   1.96  83.7   0.01 128.0   1.44  91.4
   1.96  86.2
   ;

proc glm;
      model peso=hierro hierro*hierro;
   run;
```

La salida con los datos del ajuste es la siguiente:

```
Regression con PROC GLM
   The GLM Procedure
   Number of observations    13

Dependent Variable: peso
                    Sum of
```



```
Source                  DF        Squares     Mean Square    F Value    Pr > F

Model                    2    3296.530589    1648.265295     164.68    <.0001
Error                   10     100.086334      10.008633
Corrected Total         12    3396.616923

                R-Square     Coeff Var      Root MSE     peso Mean

                0.970534      2.907348      3.163642      108.8154

Source                  DF     Type I SS    Mean Square    F Value    Pr > F

hierro                   1    3293.766690    3293.766690     329.09    <.0001
hierro*hierro            1       2.763899       2.763899       0.28    0.6107

Source                  DF    Type III SS    Mean Square    F Value    Pr > F

hierro                   1    356.7572421    356.7572421      35.64    0.0001
hierro*hierro            1      2.7638994      2.7638994       0.28    0.6107

                                     Standard
      Parameter          Estimate       Error    t Value    Pr > |t|

      Intercept        130.3199337  1.77096213      73.59      <.0001
      hierro           -26.2203900  4.39177557      -5.97      0.0001
      hierro*hierro      1.1552018  2.19828568       0.53      0.6107
```

Se observa que el ajuste global del modelo es correcto al 95% ya que el p-valor del contraste de la F es menor que 0,05 (<0,001) y el R^2 es muy alto (0,97). En cuanto a la significatividad de los coeficientes, el único término no significativo al 95% es el término cuadrático cuyo p-valor en el contraste de la T es mucho más alto que 0,05 (0,61). Esto nos sugiere que debemos eliminar el término cuadrático del ajuste y hacer el siguiente ajuste lineal:

```
proc glm;
      model peso=hierro ;
   run;
```

La salida es la siguiente:

```
                    The GLM Procedure
              Number of observations    13
                Regression con PROC GLM

                Dependent Variable: peso

                          Sum of
Source                 DF        Squares     Mean Square    F Value    Pr > F
Model                   1    3293.766690    3293.766690     352.27    <.0001
Error                  11     102.850233       9.350021
Corrected Total        12    3396.616923

                R-Square     Coeff Var      Root MSE     peso Mean

                0.969720      2.810063      3.057780      108.8154

Source                 DF     Type I SS    Mean Square    F Value    Pr > F
hierro                  1    3293.766690    3293.766690     352.27    <.0001

Source                 DF    Type III SS    Mean Square    F Value    Pr > F
hierro                  1    3293.766690    3293.766690     352.27    <.0001
```

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	129.7865993	1.40273671	92.52	<.0001
hierro	-24.0198934	1.27976715	-18.77	<.0001

Se observa que ahora el ajuste ya es correcto porque tanto los parámetros como el modelo global son altamente significativos (superior al 99%). Los p-valores de la F para la significatividad global y de la T para la significatividad de los parámetros son inferiores a una diezmilésima y el coeficiente de determinación sigue siendo muy alto (0,97).

Como segundo ejemplo consideramos un experimento que analiza los efectos de las drogas (dos antibióticos A y B y una de control F) en el tratamiento de la lepra siendo X la puntuación de leprosidad antes del tratamiento e Y después. 10 pacientes fueron seleccionados para cada tratamiento (DROGA) y se realizaron 6 medidas de la puntuación de leprosidad en cada paciente según los datos:

Droga	Pre	Pos	Droga	Pre	Pos	Droga	Pre	Pos	Droga	Pre	Pos	Droga	Pre	Pos
A	11	6	A	8	0	A	5	2	A	14	8	A	19	11
A	6	4	A	10	13	A	6	1	A	11	8	A	3	0
D	6	0	D	6	2	D	7	3	D	8	1	D	18	18
D	8	4	D	19	14	D	8	9	D	5	1	D	15	9
F	16	13	F	13	10	F	11	18	F	9	5	F	21	23
F	16	12	F	12	5	F	12	16	F	7	1	F	12	20

Con la finalidad de incrementar la precisión en la determinación del efecto de la droga después del tratamiento, se trata de ajustar los datos a un modelo del análisis de la covarianza considerando como covariable la puntuación de leprosidad antes del tratamiento (X).

Utilizaremos el siguiente programa SAS:

```
data drugtest;
      input Droga $ X Y @@;
      datalines;
   A 11  6    A  8  0    A  5  2    A 14  8    A 19 11
   A  6  4    A 10 13    A  6  1    A 11  8    A  3  0
   D  6  0    D  6  2    D  7  3    D  8  1    D 18 18
   D  8  4    D 19 14    D  8  9    D  5  1    D 15  9
   F 16 13    F 13 10    F 11 18    F  9  5    F 21 23
   F 16 12    F 12  5    F 12 16    F  7  1    F 12 20
;

   proc glm;
      class Droga;
      model Y = Droga X / solution;
      lsmeans Droga / stderr pdiff cov out=adjmeans;
   run;
   proc print data=adjmeans;
   run;
```


La salida es la siguiente:

```
                              The GLM Procedure

                          Class Level Information

                        Class        Levels    Values

                        Droga             3    A D F

                     Number of observations    30

                          Dependent Variable: Y

                                   Sum of
Source                    DF        Squares     Mean Square    F Value    Pr > F

Model                      3     871.497403      290.499134      18.10    <.0001
Error                     26     417.202597       16.046254
Corrected Total           29    1288.700000

                R-Square     Coeff Var      Root MSE        Y Mean
                0.676261      50.70604      4.005778      7.900000

Source                    DF      Type I SS     Mean Square    F Value    Pr > F

Droga                      2    293.6000000    146.8000000       9.15    0.0010
X                          1    577.8974030    577.8974030      36.01    <.0001


Source                    DF    Type III SS     Mean Square    F Value    Pr > F

Droga                      2     68.5537106     34.2768553       2.14    0.1384
X                          1    577.8974030    577.8974030      36.01    <.0001

                                          Standard
      Parameter          Estimate            Error    t Value    Pr > |t|

      Intercept      -0.434671164 B      2.47135356      -0.18      0.8617
      Droga     A    -3.446138280 B      1.88678065      -1.83      0.0793
      Droga     D    -3.337166948 B      1.85386642      -1.80      0.0835
      Droga     F     0.000000000 B       .                  .          .
      X                0.987183811      0.16449757       6.00      <.0001

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve
      the normal equations.  Terms whose estimates are followed by the letter 'B' are not
      uniquely estimable.
                              The GLM Procedure
                             Least Squares Means

                                        Standard                    LSMEAN
      Droga        Y LSMEAN                Error     Pr > |t|       Number

      A           6.7149635            1.2884943      <.0001            1
      D           6.8239348            1.2724690      <.0001            2
      F          10.1611017            1.3159234      <.0001            3

                   Least Squares Means for effect Droga
                   Pr > |t| for H0: LSMean(i)=LSMean(j)

                          Dependent Variable: Y

                i/j          1            2            3

                  1                    0.9521       0.0793
                  2       0.9521                    0.0835
                  3       0.0793       0.0835
```

```
NOTE: To ensure overall protection level, only probabilities associated with pre-planned
      comparisons should be used.
```

Obs	_NAME_	Droga	LSMEAN	STDERR	NUMBER	COV1	COV2	COV3
1	Y	A	6.7150	1.28849	1	1.66022	0.02844	-0.08403
2	Y	D	6.8239	1.27247	2	0.02844	1.61918	-0.04299
3	Y	F	10.1611	1.31592	3	-0.08403	-0.04299	1.73165

El valor Type I SS para droga (293,6) da las sumas de cuadrados entre drogas obtenidas en el análisis de la varianza *Y=droga*. El valor Type III SS para *droga* (68,5537) da las sumas de cuadrados de droga ajustadas por la covarianza. El test tipo I es altamente significativo (p=0.001), pero el test tipo III no es significativo. Esto significa que mientras existe una diferencia significativa entre las medias de las drogas, la diferencia se reduce cuando se consideran las puntuaciones en el pretratamiento (la covariable). De la tabla de parámetros estimados se puede derivar la forma del modelo de ajuste, que permitirá predecir niveles de leprosidad después del tratamiento según niveles antes del tratamiento y droga suministrada. El modelo es el siguiente:

$$Y=\begin{cases}(-0{,}435+-3.446)+0{,}987X & si \quad Droga=A\\ (-0{,}435+-3.337)+0{,}987X & si \quad Droga=D\\ -0{,}435 \qquad\qquad +0{,}987X & si \quad Droga=F\end{cases}$$

COMPONENTES DE LA VARIANZA EN SAS: PROCEDIMIENTO VARCOMP

El procedimiento VARCOMP estima las componentes de la varianza en un modelo lineal general. Su sintaxis es la siguiente:

PROC VARCOM opciones;
CLASS variables;
MODEL dependientes = efectos / opción;
BY variables;

Las opciones de PROC VARCOMP son DATA=conjunto de datos de entrada, EPSILON=nº para el criterio de convergencia, MAXITER=nº máximo de iteraciones y METHOD=TYPE1 | MIVQUE0 | ML | REML para especificar el método de convergencia. La opción de MODEL es FIXED=n que especifica que los *n* primeros efectos del modelo son fijos.

Como primer ejemplo consideremos las variables de clasificación *a* y *b* y la variable dependiente Y. Sea *a* un efecto fijo y *b* y *a*b* efectos aleatorios. Vamos a estimar las componentes de la varianza de este modelo no balanceado (tamaños distintos en las celdas) por los cuatro métodos de estimación.

```
data a;
      input a b y @@;
      datalines;
```



```
   1 1 237   1 1 254   1 1 246   1 2 178   1 2 179
   2 1 208   2 1 178   2 1 187   2 2 146   2 2 145   2 2 141
   3 1 186   3 1 183   3 2 142   3 2 125   3 2 136
;
     proc varcomp method=type1;
      class a b;
      model y=a|b / fixed=1;
   run;
     proc varcomp method=mivque0;
      class a b;
      model y=a|b / fixed=1;
   run;
     proc varcomp method=ml;
      class a b;
      model y=a|b / fixed=1;
   run;
```

La salida es la siguiente:

```
                    Variance Components Estimation Procedure

                          Class Level Information

                      Class         Levels    Values

                      a                  3    1 2 3
                      b                  2    1 2

                       Number of observations    16

                         Dependent Variable:    y

                       Type 1 Analysis of Variance

                              Sum of
Source              DF       Squares   Mean Square  Expected Mean Square

a                    2         11736   5868.218750  Var(Error) + 2.725 Var(a*b) + 0.1 Var(b) + Q(a)
b                    1         11448         11448  Var(Error) + 2.6308 Var(a*b) + 7.8 Var(b)
a*b                  2    299.041026    149.520513  Var(Error) + 2.5846 Var(a*b)
Error               10    786.333333     78.633333  Var(Error)
Corrected Total     15         24270             .  .

                            Type 1 Estimates
                       Variance Component        Estimate

                       Var(b)                      1448.4
                       Var(a*b)                   27.42659
                       Var(Error)                 78.63333

                    Variance Components Estimation Procedure
                          Class Level Information

                      Class         Levels    Values

                      a                  3    1 2 3
                      b                  2    1 2
                       Number of observations    16
                           MIVQUE(0) SSQ Matrix
  Source                  b                 a*b              Error                  y

  b                 60.84000           20.52000            7.80000            89295.4
  a*b               20.52000           20.52000            7.80000            30181.3
  Error              7.80000            7.80000           13.00000            12533.5
```

```
                    MIVQUE(0) Estimates

          Variance Component              y

          Var(b)                     1466.1
          Var(a*b)                -35.49170
          Var(Error)              105.73660

       Variance Components Estimation Procedure


                 Class Level Information

            Class          Levels    Values

            a                   3    1 2 3
            b                   2    1 2

             Number of observations     16
               Dependent Variable:      y
             Maximum Likelihood Iterations

Iteration        Objective          Var(b)          Var(a*b)        Var(Error)

        0     78.3850371200       1031.49070                 0     74.3909717935
        1     78.2637043807    732.3606453635                0     77.4011688154
        2     78.2635471161    723.6867470850                0     77.5301774839
        3     78.2635471152    723.6658365289                0     77.5304926877

   Convergence criteria met.

                    Maximum Likelihood
                        Estimates

                 Variance
                 Component        Estimate

                 Var(b)          723.66584
                 Var(a*b)                0
                 Var(Error)       77.53049

       Asymptotic Covariance Matrix of Estimates

                     Var(b)        Var(a*b)       Var(Error)

Var(b)             537826.1               0        -107.33905
Var(a*b)                  0               0                 0
Var(Error)        -107.33905               0         858.71104
```

SAS Y LOS MODELOS MIXTOS. PROC MIXED

El PROC MIXED de SAS permite trabajar con gran variedad de modelos lineales mixtos que se utilizan habitualmente en el análisis estadístico. Este procedimiento incluye:

- Estructuras de la covarianza diagonal y general, incluyendo componentes de la varianza, AR(1), Toeplitz, espacial, lineal general y factor analitico.
- Sintaxis tipo GLM, usando sentencias MODEL, RANDOM y REPEATED para especificación de modelos y usando sentencias CONTRAST, ESTIMATE, y LSMEANS para especificar inferencia en los modelos.
- Errores estándar apropiados para todas las combinaciones lineales estimables de efectos fijos y aleatorios con sus correspondientes tests de la F y la t.

- Efectos de grupo que permiten trabajar con bloques y heterogeneidad.
- Métodos de estimación REML y ML con el algoritmo de Newton-Raphson.
- Capacidad para trabajar con datos no balanceados.

La sintaxis general del PROC MIXED es la siguiente:

PROC MIXED < opciones > ;
BY variables ;
CLASS variables ;
ID variables ;
MODEL dependentiente = < efectos fijos > < / opciones > ;
RANDOM efectos aleatorios < / opciones > ;
REPEATED < efectos repetidos > < / opciones > ;
PARMS (lista de variables) ... < / opciones > ;
PRIOR < distribución > < / opciones > ;
CONTRAST 'etiqueta' < valores de efectos fijos ... >
< \| valores de efectos aleatorios ... >, ... < / opciones >
;
ESTIMATE 'etiqueta' < valores de efectos fijos ... >
< \| valores de efectos aleatorios ... >< / opciones > ;
LSMEANS efectos fijos < / opciones > ;
WEIGHT variable ;

Tomamos como ejemplo un diseño en parcelas divididas que es uno de los modelos mixtos más comunes. El ***diseño split-splot o diseño en parcelas divididas*** es una extensión del diseño de bloques al azar cuyo origen es el análisis agrario. El concepto *split-splot* se refiere a una parcela de terreno que se subdivide (*split*) en varias porciones (*splot*). Estos diseños también se aplican a la investigación educativa. Este modelo se utiliza cuando se combinan dos factores A y B y se obtienen réplicas organizadas en bloques. El factor bloque C tiene un efecto principal, pero no interacciona con A y B a la vez. En este diseño se comparan a tratamientos (factor A) que se asignan aleatoriamente en b bloques o parcelas (factor B), a razón de a tratamientos por bloque. Se divide cada una de las ab parcelas y se asignan al azar c subtratamientos a estas divisiones (factor C). Se supone que actúan los efectos principales A,B,C, la interacción AxC y la interacción AxB. La interacción entre A y los bloques es debida a que éstos no pueden considerarse completamente homogéneos. Sin embargo, se supone que cada una de las ab parcelas dentro de los bloques son homogéneas para que los subtratamientos C no interaccionen con los bloques. El modelo podría expresarse en la forma:

$$X_{ijkl} = \mu+A_i+B_j+C_k+AB_{ij}+AC_{ik}+\varepsilon_{ijkl} \quad i = 1..a, \quad j = 1..b, \quad k = 1...c, \quad l = 1..n_{ijk}$$

Como ejemplo consideramos un modelo mixto cuya variable respuesta Y recoge las mediciones de un experimento existiendo un efecto fijo A y dos efectos aleatorios B y A*B. Se estimará el modelo graficando la superficie de verosimilitud mediante la siguiente sintaxis SAS:

```
data hh;
  input a b y @@;
  datalines;
1 1 237   1 1 254    1 1 246
1 2 178   1 2 179
2 1 208   2 1 178    2 1 187
2 2 146   2 2 145    2 2 141
3 1 186   3 1 183
3 2 142   3 2 125    3 2 136
;
ods output ParmSearch=parms;
proc mixed data=hh asycov mmeq mmeqsol covtest;
   class a b;
   model y = a / outp=predicted;
   random b a*b;
   lsmeans a;
   parms (17 to 20 by .1) (.3 to .4 by .005) (1.0);
run;
proc print data=predicted;
run;
```

La salida es la siguiente:

Procedimiento Mixed
Información del modelo

Conjunto de datos	WORK.HH
Variable dependiente	y
Estructura de covarianza	Variance Components
Método de estimación	REML
Método de varianza del residual	Perfil
Método SE de efectos fijos	Basado en el modelo
Método de grados de libertad	Contención

Información de nivel de clase

Clase	Niveles	Valores
a	3	1 2 3
b	2	1 2

Dimensiones

Parámetros de covarianza	3
Columnas en X	4
Columnas en Z	8
Asuntos	1
Obs máx por asunto	16

Número de observaciones

Número de observaciones leídas	16
Número de observaciones usadas	16
Número de observaciones no usada	0

Historia de iteración

Iteración	Evaluaciones	-2 Res Log Like	Criterio
1	2	104.93416367	0.00000000

Criterio de convergencia cumplido.

Estimadores de parámetro de covarianza

Parm Cov	Estimador	Error estándar	Valor Z	Pr > Z
b	1464.36	2098.01	0.70	0.2426
a*b	26.9581	59.6570	0.45	0.3257
Residual	78.8426	35.3512	2.23	0.0129

Matriz de covarianza asintótica de los estimadores

Fila	Parm Cov	CovP1	CovP2	CovP3
1	b	4401640	1.2831	-273.32
2	a*b	1.2831	3558.96	-502.84
3	Residual	-273.32	-502.84	1249.71

Estadísticos de ajuste

Verosimilitud -2 Res Log	104.9
AIC (mejor más pequeño)	110.9
AICC (mejor más pequeño)	113.6
BIC (mejor más pequeño)	107.0

Test del ratio de verosimilitud del modelo PARMS

DF	Chi-cuadrado	Pr > ChiSq
2	0.00	1.0000

Ecuaciones de modelo mixto

Fila	Efecto	a	b	Col1	Col2	Col3	Col4	Col5	Col6	Col7
1	Intercept			0.2029	0.06342	0.07610	0.06342	0.1015	0.1015	0.03805
2	a	1		0.06342	0.06342			0.03805	0.02537	0.03805
3	a	2		0.07610		0.07610		0.03805	0.03805	
4	a	3		0.06342			0.06342	0.02537	0.03805	
5	b		1	0.1015	0.03805	0.03805	0.02537	0.1022		0.03805
6	b		2	0.1015	0.02537	0.03805	0.03805		0.1022	
7	a*b	1	1	0.03805	0.03805			0.03805		0.07515
8	a*b	1	2	0.02537	0.02537				0.02537	
9	a*b	2	1	0.03805		0.03805		0.03805		
10	a*b	2	2	0.03805		0.03805			0.03805	
11	a*b	3	1	0.02537			0.02537	0.02537		
12	a*b	3	2	0.03805			0.03805		0.03805	

Ecuaciones de modelo mixto

Fila	Col8	Col9	Col10	Col11	Col12	Col13
1	0.02537	0.03805	0.03805	0.02537	0.03805	36.4143
2	0.02537					13.8757
3		0.03805	0.03805			12.7469
4				0.02537	0.03805	9.7917
5		0.03805		0.02537		21.2956
6	0.02537		0.03805		0.03805	15.1187
7						9.3477
8	0.06246					4.5280
9		0.07515				7.2676
10			0.07515			5.4793
11				0.06246		4.6802
12					0.07515	5.1115

Solución de ecuaciones de modelo mixto

Fila	Efecto	a	b	Col1	Col2	Col3	Col4	Col5	Col6	Col7
1	Intercept			761.84	-29.7718	-29.6578		-731.14	-733.22	-0.4680
2	a	1		-29.7718	59.5436	29.7718		-2.0764	2.0764	-14.0239
3	a	2		-29.6578	29.7718	56.2773		-1.0382	1.0382	0.4680
4	a	3								
5	b		1	-731.14	-2.0764	-1.0382		741.63	722.73	-4.2598
6	b		2	-733.22	2.0764	1.0382		722.73	741.63	4.2598
7	a*b	1	1	-0.4680	-14.0239	0.4680		-4.2598	4.2598	22.8027
8	a*b	1	2	0.4680	-12.9342	-0.4680		4.2598	-4.2598	4.1555
9	a*b	2	1	-0.5257	1.0514	-12.9534		-4.7855	4.7855	2.1570
10	a*b	2	2	0.5257	-1.0514	-14.0048		4.7855	-4.7855	-2.1570
11	a*b	3	1	-12.4663	12.9342	12.4663		-4.2598	4.2598	1.9200
12	a*b	3	2	-14.4918	14.0239	14.4918		4.2598	-4.2598	-1.9200

Solución de ecuaciones de modelo mixto

Fila	Col8	Col9	Col10	Col11	Col12	Col13
1	0.4680	-0.5257	0.5257	-12.4663	-14.4918	159.61
2	-12.9342	1.0514	-1.0514	12.9342	14.0239	53.2049
3	-0.4680	-12.9534	-14.0048	12.4663	14.4918	7.8856
4						
5	4.2598	-4.7855	4.7855	-4.2598	4.2598	26.8837
6	-4.2598	4.7855	-4.7855	4.2598	-4.2598	-26.8837
7	4.1555	2.1570	-2.1570	1.9200	-1.9200	3.0198
8	22.8027	-2.1570	2.1570	-1.9200	1.9200	-3.0198
9	-2.1570	22.5560	4.4021	2.1570	-2.1570	-1.7134
10	2.1570	4.4021	22.5560	-2.1570	2.1570	1.7134
11	-1.9200	2.1570	-2.1570	22.8027	4.1555	-0.8115
12	1.9200	-2.1570	2.1570	4.1555	22.8027	0.8115

Tests de tipo 3 de efectos fijos

Efecto	Num DF	Den DF	F-Valor	Pr > F
a	2	2	28.00	0.0345

Procedimiento Mixed

Medias de mínimos cuadrados

Efecto	a	Estimador	Error estándar	DF	Valor t	Pr > |t|
a	1	212.82	27.6014	2	7.71	0.0164
a	2	167.50	27.5463	2	6.08	0.0260
a	3	159.61	27.6014	2	5.78	0.0286

Valores predichos

Obs	a	b	y	Pred	StdErr Pred	DF	Alpha	Lower	Upper	Resid
1	1	1	237	242.723	4.72563	10	0.05	232.193	253.252	-5.7228
2	1	1	254	242.723	4.72563	10	0.05	232.193	253.252	11.2772
3	1	1	246	242.723	4.72563	10	0.05	232.193	253.252	3.2772
4	1	2	178	182.916	5.52589	10	0.05	170.603	195.228	-4.9159
5	1	2	179	182.916	5.52589	10	0.05	170.603	195.228	-3.9159
6	2	1	208	192.670	4.70076	10	0.05	182.196	203.144	15.3297
7	2	1	178	192.670	4.70076	10	0.05	182.196	203.144	-14.6703
8	2	1	187	192.670	4.70076	10	0.05	182.196	203.144	-5.6703
9	2	2	146	142.330	4.70076	10	0.05	131.856	152.804	3.6703
10	2	2	145	142.330	4.70076	10	0.05	131.856	152.804	2.6703
11	2	2	141	142.330	4.70076	10	0.05	131.856	152.804	-1.3297
12	3	1	186	185.687	5.52589	10	0.05	173.374	197.999	0.3134
13	3	1	183	185.687	5.52589	10	0.05	173.374	197.999	-2.6866
14	3	2	142	133.542	4.72563	10	0.05	123.013	144.072	8.4578
15	3	2	125	133.542	4.72563	10	0.05	123.013	144.072	-8.5422
16	3	2	136	133.542	4.72563	10	0.05	123.013	144.072	2.4578

El gráfico de la superficie de verosimilitud se presenta en la Figura 8-70.

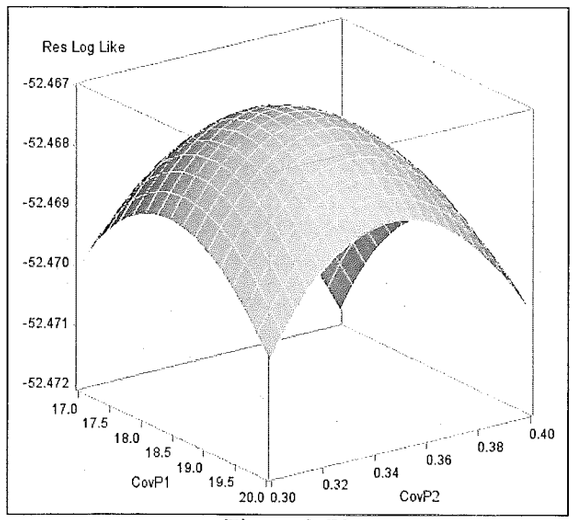


Figura 8-70

STATA Y EL ANÁLISIS DE LA VARIANZA-COVARIANZA, EL MODELO GLM Y LOS MODELOS MIXTOS

El comando *anova* de STATA permite realizar análisis de la varianza y la covarianza simples. Como ejemplo se ejecuta un experimento variando la cantidad de fertilizante utilizado para el crecimiento de manzanos. Contrastamos cuatro concentraciones usando cada concentración en tres fincas de doce manzanos cada una. Después de un año medimos el peso medio de la fruta. Mediante el análisis de la varianza contrastamos si el peso medio difiere para los diferentes tratamientos.

```
. use http://www.stata-press.com/data/r9/apple
(Apple trees)

. anova weight treatment

                       Number of obs =      10     R-squared     =  0.9147
                       Root MSE      = 9.07002     Adj R-squared =  0.8721

           Source |  Partial SS    df       MS           F     Prob > F
-----------+----------------------------------------------------------
            Model |  5295.54433     3  1765.18144       21.46     0.0013
                  |
        treatment |  5295.54433     3  1765.18144       21.46     0.0013
                  |
         Residual |  493.591667     6  82.2652778
-----------+----------------------------------------------------------
            Total |    5789.136     9  643.237333
```

Los p-valores del resultado aceptan las diferencias significativas entre los diferentes tratamientos realizados con las cuatro concentraciones.

Podemos obtener los coeficientes de ajuste del modelo ANOVA como un modelo de regresión.

```
. anova, regress

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  3,     6) =   21.46
       Model |  5295.54433     3  1765.18144           Prob > F      =  0.0013
    Residual |  493.591667     6  82.2652778           R-squared     =  0.9147
-------------+------------------------------           Adj R-squared =  0.8721
       Total |    5789.136     9  643.237333           Root MSE      =    9.07

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |        77.5   6.413473    12.08   0.000     61.8068     93.1932
   treatment |
           1 |        34.4   8.279758     4.15   0.006     14.14016    54.65984
           2 |   -24.76667   8.279758    -2.99   0.024    -45.02651   -4.506828
           3 |        1.15   9.070021     0.13   0.903    -21.04354    23.34354
           4 |   (dropped)
------------------------------------------------------------------------------
```

Se puede realizar ***análisis de la varianza simple con varios factores*** incluyendo interacciones de diferentes órdenes entre las variables.


```
. use http://www.stata-press.com/data/r9/manuf
(manufacturing process data)

. anova yield temp chem temp*chem meth temp*meth chem*meth temp*chem*meth

                       Number of obs =      36     R-squared     =  0.5474
                       Root MSE      = 2.62996     Adj R-squared =  0.3399

                 Source |  Partial SS    df       MS           F     Prob > F
----------------------+----------------------------------------------------
                  Model |     200.75    11    18.25        2.64     0.0227
                        |
            temperature |       30.5     2    15.25        2.20     0.1321
                chemical |     12.25     1    12.25        1.77     0.1958
   temperature*chemical |      24.5     2    12.25        1.77     0.1917
                 method |     42.25     1    42.25        6.11     0.0209
     temperature*method |      87.5     2    43.75        6.33     0.0062
        chemical*method |       .25     1      .25        0.04     0.8508
 temperature*chemical*  |
                 method |       3.5     2      1.75        0.25     0.7785
                        |
               Residual |       166    24  6.91666667
----------------------+----------------------------------------------------
                  Total |     366.75    35  10.4785714
```

También se puede realizar ***análisis de la covarianza simple*** considerando variables explicativas categóricas y numéricas mezcladas, en cuyo caso es necesario especificar las variables numéricas mediante *continuous*(*variable*1 *variable*2...).

```
. anova drate region age, continuous(age)

                       Number of obs =      50     R-squared     =  0.7203
                       Root MSE      = 7.21483     Adj R-squared =  0.6954

                 Source |  Partial SS    df       MS           F     Prob > F
----------------------+----------------------------------------------------
                  Model |  6032.08254     4  1508.02064       28.97     0.0000
                        |
                 region |  1645.66228     3  548.554092       10.54     0.0000
                    age |  1630.46662     1  1630.46662       31.32     0.0000
                        |
               Residual |  2342.41746    45  52.0537213
----------------------+----------------------------------------------------
                  Total |     8374.5    49  170.908163

. anova, regress

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  4,    45) =   28.97
       Model |  6032.08254     4  1508.02064           Prob > F      =  0.0000
    Residual |  2342.41746    45  52.0537213           R-squared     =  0.7203
-------------+------------------------------           Adj R-squared =  0.6954
       Total |     8374.5    49  170.908163           Root MSE      =  7.2148

------------------------------------------------------------------------------
       drate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -41.97859   19.88906    -2.11   0.040    -82.03722   -1.919959
      region |
           1 |   13.37578   3.723447     3.59   0.001     5.876377    20.87519
           2 |   15.16831   3.022157     5.02   0.000     9.081374    21.25525
           3 |   14.07378   2.865756     4.91   0.000     8.301846     19.8457
           4 |  (dropped)
         age |   3.922947   .7009425     5.60   0.000     2.511177    5.334718
------------------------------------------------------------------------------
```

STATA realiza el análisis de la varianza y la covarianza múltiples mediante el comando *manova*. A continuación se presenta un ejemplo de ***análisis de la varianza múltiple***.

```
. use http://www.stata-press.com/data/r9/manuf
(manufacturing process data)

. manova yield method = temp chem

                    Number of obs =       36

                    W = Wilks' lambda       L = Lawley-Hotelling trace
                    P = Pillai's trace      R = Roy's largest root

        Source |  Statistic     df   F(df1,     df2) =    F   Prob>F
-------------+-----------------------------------------------------
         Model | W   0.8683      3      6.0      62.0      0.76 0.6069 e
               | P   0.1317             6.0      64.0      0.75 0.6100 a
               | L   0.1517             6.0      60.0      0.76 0.6052 a
               | R   0.1517             3.0      32.0      1.62 0.2045 u
               |-----------------------------------------------------
      Residual |                32
-------------+-----------------------------------------------------
   temperature | W   0.9023      2      4.0      62.0      0.82 0.5190 e
               | P   0.0977             4.0      64.0      0.82 0.5162 a
               | L   0.1083             4.0      60.0      0.81 0.5225 a
               | R   0.1083             2.0      32.0      1.73 0.1931 u
               |-----------------------------------------------------
       chemical | W   0.9583      1      2.0      31.0      0.67 0.5170 e
               | P   0.0417             2.0      31.0      0.67 0.5170 e
               | L   0.0435             2.0      31.0      0.67 0.5170 e
               | R   0.0435             2.0      31.0      0.67 0.5170 e
               |-----------------------------------------------------
      Residual |                32
-------------+-----------------------------------------------------
         Total |                35
-----------------------------------------------------------------------
                    e = exact, a = approximate, u = upper bound on F
```

A continuación se presenta un ejemplo de ***análisis de la covarianza múltiple***.

```
. manova yield method = temp chem, continuous(temp)

                    Number of obs =       36

                    W = Wilks' lambda       L = Lawley-Hotelling trace
                    P = Pillai's trace      R = Roy's largest root
        Source |  Statistic     df   F(df1,     df2) =    F   Prob>F
-------------+-----------------------------------------------------
         Model | W   0.8686      2      4.0      64.0      1.17 0.3337 e
               | P   0.1314             4.0      66.0      1.16 0.3366 a
               | L   0.1512             4.0      62.0      1.17 0.3319 a
               | R   0.1512             2.0      33.0      2.50 0.0979 u
               |-----------------------------------------------------
      Residual |                33
-------------+-----------------------------------------------------
   temperature | W   0.9027      1      2.0      32.0      1.72 0.1945 e
               | P   0.0973             2.0      32.0      1.72 0.1945 e
               | L   0.1078             2.0      32.0      1.72 0.1945 e
               | R   0.1078             2.0      32.0      1.72 0.1945 e
```



```
               |-----------------------------------------------------
       chemical | W   0.9584      1      2.0      32.0      0.70 0.5063 e
               | P   0.0416             2.0      32.0      0.70 0.5063 e
               | L   0.0435             2.0      32.0      0.70 0.5063 e
               | R   0.0435             2.0      32.0      0.70 0.5063 e
               |-----------------------------------------------------
      Residual |                33
-------------+-----------------------------------------------------
         Total |                35
-----------------------------------------------------------------------
                    e = exact, a = approximate, u = upper bound on F
```

STATA permite ajustar modelos lineales generalizados mediante el comando *glm*, pudiendo incluir en el modelo variables normales, binomiales, logísticas, binomiales negativas, gamma y de Poisson. A continuación se presenta un ejemplo.

```
. use http://www.stata-press.com/data/r9/lbw
(Hosmer & Lemeshow data)

. xi: glm low age lwt i.race smoke ptl ht ui, f(bin) l(logit)
i.race            _Irace_1-3          (naturally coded; _Irace_1 omitted)

Iteration 0:   log likelihood =  -101.0213
Iteration 1:   log likelihood = -100.72519
Iteration 2:   log likelihood =   -100.724
Iteration 3:   log likelihood =   -100.724

Generalized linear models                          No. of obs      =       189
Optimization     : ML                              Residual df     =       180
                                                   Scale parameter =         1
Deviance         =  201.4479911                    (1/df) Deviance =  1.119156
Pearson          =  182.0233425                    (1/df) Pearson  =  1.011241

Variance function: V(u) = u*(1-u)                  [Bernoulli]
Link function    : g(u) = ln(u/(1-u))              [Logit]

                                                   AIC             =    1.1611
Log likelihood   = -100.7239956                    BIC             = -742.0665
------------------------------------------------------------------------------
                 |                 OIM
             low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
             age |  -.0271003   .0364504    -0.74   0.457    -.0985418    .0443412
             lwt |  -.0151508   .0069259    -2.19   0.029    -.0287253   -.0015763
         _Irace_2 |   1.262647   .5264101     2.40   0.016     .2309024    2.294392
         _Irace_3 |   .8620792   .4391532     1.96   0.050     .0013548    1.722804
           smoke |   .9233448   .4008266     2.30   0.021      .137739    1.708951
             ptl |   .5418366    .346249     1.56   0.118     -.136799    1.220472
              ht |   1.832518   .6916292     2.65   0.008     .4769494    3.188086
              ui |   .7585135   .4593768     1.65   0.099    -.1418484    1.658875
           _cons |   .4612239    1.20459     0.38   0.702    -1.899729    2.822176
------------------------------------------------------------------------------
```

STATA permite ajustar modelos lineales de efectos mixtos multinivel a través del comando *xtmixed*. A continuación se presenta un ejemplo.

```
. xi: xtmixed low age lwt i.race smoke ptl ht ui
i.race            _Irace_1-3          (naturally coded; _Irace_1 omitted)
```

```
Mixed-effects REML regression                   Number of obs      =       189

                                                Wald chi2(8)       =     35.19
Log restricted-likelihood = -127.95413          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0034688   .0063194    -0.55   0.583    -.0158546     .008917
         lwt |  -.0025213   .0011532    -2.19   0.029    -.0047816    -.000261
    _Irace_2 |   .2214043   .1001543     2.21   0.027     .0251055     .417703
    _Irace_3 |   .1436247   .0765303     1.88   0.061    -.0063719    .2936213
       smoke |   .1595568   .0710842     2.24   0.025     .0202343    .2988794
         ptl |   .1153871     .06806     1.70   0.090     -.018008    .2487822
          ht |   .3635326    .134455     2.70   0.007     .1000057    .6270596
          ui |   .1560515   .0927102     1.68   0.092    -.0256572    .3377602
       _cons |   .5074597   .2085242     2.43   0.015     .0987598    .9161596
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
               sd(Residual) |   .4342674   .0223363      .3926234    .4803285
------------------------------------------------------------------------------
```

Ejercicio 8-1. A partir de los datos del archivo empleados.sav se trata de comprobar si los grupos definidos por la categoría laboral (catlab) y los grupos definidos por la clasificación étnica (minoría), difieren en su salario actual (salario). En caso de que los grupos difieran comprobar si las diferencias observadas en salario entre los distintos subgrupos definidos por categoría laboral y clasificación étnica, se mantienen al controlar el efecto de las variables experiencia previa (expprev) y tiempo desde el contrato (tiempemp).

Para responder a la primera cuestión se realiza un análisis univariante de la varianza bifactorial donde la variable dependiente es el salario (*salario*) y los dos factores son la categoría laboral (*catlab*) y la clasificación étnica (minoría). Para realizar el análisis univariante de la varianza bifactorial, elija en los menús *Analizar* → *Modelo lineal general* → *Univariante* (Figura 8-71), seleccione la variable dependiente salario y seleccione catlab y minoría como *Factores fijos* (Figura 8-72).

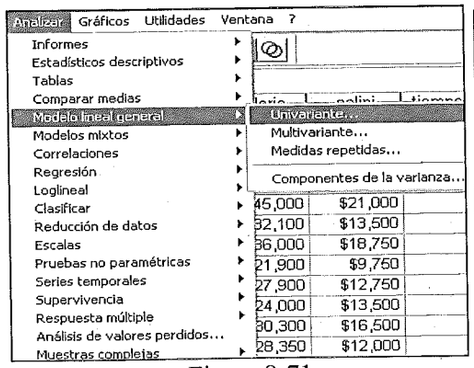


Figura 8-71

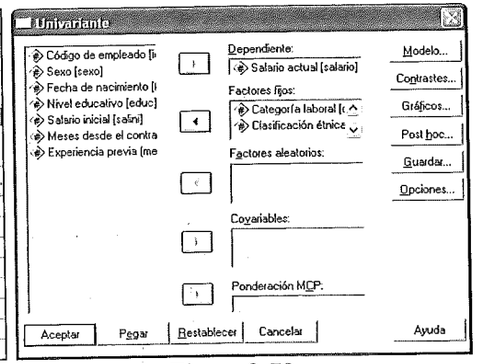


Figura 8-72

El botón *Gráficos* se rellena como se indica en la Figura 8-73 con la finalidad de definir los gráficos de perfil (gráficos de interacción) que sirven para comparar las medias marginales en el modelo. El botón *Post hoc* se cumplimenta como se indica en la Figura 8-74 para obtener los contrastes post hoc para comparar las medias observadas entre sí. El botón *Opciones* se cumplimenta como se indica en la Figura 8-75 para seleccionar estadísticos adicionales, estimaciones puntuales y por intervalos en el modelo.

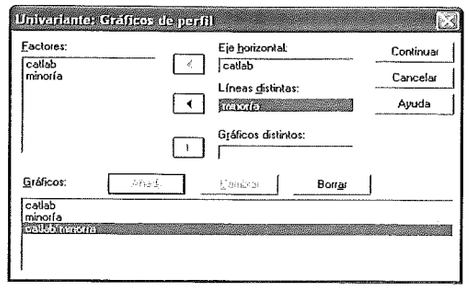


Figura 8-73

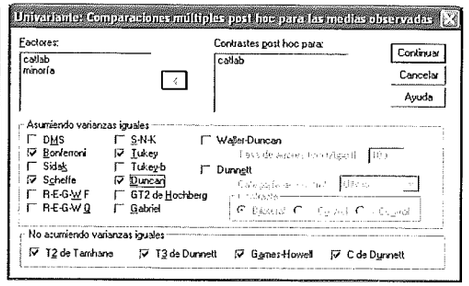


Figura 8-74

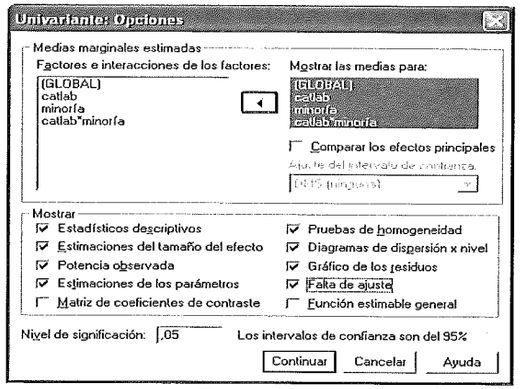


Figura 8-75

Al pulsar *Aceptar* en la Figura 8-72, se obtiene la salida. En la Figura 8-76 vemos el *Contraste de Levenne*, cuyo p-valor muy pequeño permite rechazar la hipótesis de igualdad de varianzas de la variable dependiente en el conjunto de poblaciones (subgrupos) definidas por la combinación de factores. También se muestra la tabla ANOVA o *Prueba de los efectos inter-sujetos* con las fuentes de variación, las sumas de cuadrados, los grados de libertad, los cuadrados medios o medias cuadráticas, los estadísticos F y los niveles de significación asociados a cada estadístico F. La fila *Modelo corregido* se refiere a todos los efectos del modelo a la vez y su p-valor pequeño indica que el modelo explica una parte significativa de la variación observada en la variable dependiente *salario*. El valor de R^2 = 0,655 indica que los tres efectos incluidos en el modelo (*catlab, minoría* y su interacción) explican el 65,5% de la varianza de la variable dependiente *salario*.

La fila *Intersección* se refiere a la constante del modelo y su p-valor pequeño indica su alta significatividad. Las filas siguientes recogen los efectos principales y la interacción. El p-valor pequeño relativo a la variable *catlab* indica que los grupos definidos por esta variable en la variable dependiente *salario* poseen salarios medios significativamente diferentes. Sin embargo, el p-valor mayor que 0,05 relativo a la variable *minoría* indica que los grupos definidos por esta variable en la variable dependiente *salario* poseen salarios medios significativamente iguales. El efecto de la interacción también resulta significativo determinando grupos de salarios distintos. La fila *Error* recoge la suma de cuadrados y la media cuadrática del error, la fila *Total* recoge la suma de cuadrados de la variable dependiente y la fila *Total corregida* recoge la variación debida a cada efecto más la variación del error (variación total).

Contraste de Levene sobre la igualdad de las varianzas error[a]

Variable dependiente: Salario actual

F	gl1	gl2	Significación
24,720	5	468	,000

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intercept+catlab+minoría+catlab * minoría

Pruebas de los efectos inter-sujetos

Variable dependiente: Salario actual

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación	Eta al cuadrado parcial	Parámetro de no centralidad	Potencia observada[a]
Modelo corregido	9,034E+10[b]	5	1,81E+10	177,742	,000	,655	888,708	1,000
Intersección	1,537E+11	1	1,54E+11	1511,773	,000	,764	1511,773	1,000
catlab	2,596E+10	2	1,30E+10	127,699	,000	,353	255,398	1,000
minoría	237964814	1	2,38E+08	2,341	,127	,005	2,341	,333
catlab * minoría	788578413	2	3,94E+08	3,879	,021	,016	7,757	,700
Error	4,757E+10	468	1,02E+08					
Total	6,995E+11	474						
Total corregida	1,379E+11	473						

a. Calculado con alfa = ,05

b. R cuadrado = ,655 (R cuadrado corregida = ,651)

Figura 8-76

La Figura 8-77 muestra las estimaciones de los parámetros del modelo que permiten obtener las medias que el modelo estima para cada nivel o combinación de niveles y que se calculan combinado los parámetros involucrados en la obtención de cada media. Por ejemplo, la estimación de la media de los administrativos blancos (*catlab* = 1 y *minoría* = 1) se obtiene sumando las estimaciones para la constante (76037,5), para *catlab* = 1 (-49793,2), para *minoría* = 1 (-12662,7) y para *catlab*=1**minoría*=1 (14759,522). Se obtiene así un salario medio de 28341,09, que es el valor que se observa para administrativos blancos en la tabla de estadísticos descriptivos de la Figura 8-78 que muestra media, desviación típica y tamaño de cada nivel y combinación de niveles para cada factor. La Figura 8-79 presenta las medias marginales estimadas para cada factor e interacciones (incluido el total).

Estimaciones de los parámetros

Variable dependiente: Salario actual

Parámetro	B	Error típ.	t	Significación	Intervalo de confianza al 95%. Límite inferior	Límite superior	Eta al cuadrado parcial	Parámetro de no centralidad	Potencia observada[a]
Intersección	76037,500	5041,212	15,083	,000	66131,287	85943,713	,327	15,083	1,000
[catlab=1]	-49793,2	5155,800	-9,658	,000	-59924,630	-39661,9	,166	9,658	1,000
[catlab=2]	-45356,7	5764,847	-7,868	,000	-56684,919	-34028,5	,117	7,868	1,000
[catlab=3]	0[b]	.	.	.	.	.	.	.	.
[minoría=1]	-12662,7	5165,705	-2,451	,015	-22813,535	-2511,840	,013	2,451	,687
[minoría=2]	0[b]	.	.	.	.	.	.	.	.
[catlab=1] * [minoría=1]	14759,522	5312,370	2,778	,006	4320,470	25198,573	,016	2,778	,792
[catlab=1] * [minoría=2]	0[b]	.	.	.	.	.	.	.	.
[catlab=2] * [minoría=1]	13160,490	6462,603	2,036	,042	461,178	25859,801	,009	2,036	,529
[catlab=2] * [minoría=2]	0[b]	.	.	.	.	.	.	.	.
[catlab=3] * [minoría=1]	0[b]	.	.	.	.	.	.	.	.
[catlab=3] * [minoría=2]	0[b]	.	.	.	.	.	.	.	.

a. Calculado con alfa = ,05

b. Al parámetro se le ha asignado el valor cero porque es redundante.

Figura 8-77

Factores inter-sujetos

		Etiqueta del valor	N
Categoría laboral	1	Administrativo	363
	2	Seguridad	27
	3	Directivo	84
Clasificación étnica	1	Sí	370
	2		104

Estadísticos descriptivos

Variable dependiente: Salario actual

Categoría laboral	Clasificación étnica	Media	Desv. típ.	N
Administrativo	Sí	28341,09	7994,659	276
	2	26244,25	5772,874	87
	Total	27838,54	7567,995	363
Seguridad	Sí	31178,57	1658,743	14
	2	30680,77	2562,920	13
	Total	30938,89	2114,616	27
Directivo	Sí	63374,81	18164,043	80
	2	76037,50	17821,961	4
	Total	63977,80	18244,776	84
Total	Sí	36023,31	18044,096	370
	2	28713,94	11421,638	104
	Total	34419,57	17075,661	474

Figura 8-78

Medias marginales estimadas

1. Media global

Variable dependiente: Salario actual

Media	Error típ.	Intervalo de confianza al 95%. Límite inferior	Límite superior
42642,832	1096,738	40487,692	44797,972

2. Categoría laboral

Variable dependiente: Salario actual

Categoría laboral	Media	Error típ.	Intervalo de confianza al 95%. Límite inferior	Límite superior
Administrativo	27292,670	619,832	26074,671	28510,669
Seguridad	30929,670	1941,696	27114,149	34745,191
Directivo	69706,156	2582,853	64630,732	74781,580

3. Clasificación étnica

Variable dependiente: Salario actual

Clasificación étnica	Media	Error típ.	Intervalo de confianza al 95%. Límite inferior	Límite superior
Sí	40964,824	994,434	39010,714	42918,933

Figura 8-79

La Figura 8-80 muestra las comparaciones múltiples *post hoc* de las medias de los subgrupos delimitados en el salario por los valores de los factores y sus combinaciones. Se observa significatividad para todas las diferencias de medias según los contrastes que suponen en varianzas desiguales (Tamhane, *T*3 de Dunnet, etc.) ya que los p-valores son muy pequeños, pero no se observa significatividad de todas las diferencias de medias para los contrastes que suponen varianzas iguales (Tukey, Scheffe, Bonferroni, etc.). Por ejemplo, de acuerdo con el test de Tukey, el grupo *Administrativos* no difiere del grupo *Agentes de seguridad* (p-valor = 0,277 > 0,05), pero los dos difieren del grupo de directivos (p-valor = 0,000). No obstante debemos quedarnos con los tests que suponen diferencia significativas de varianzas, hipótesis que ya fue probada mediante la *prueba de Levenne de igualdad de varianzas* (Figura 8-76).

Comparaciones múltiples

Variable dependiente: Salario actual

	(I) Categoría laboral	(J) Categoría laboral	Diferencia entre medias (I-J)	Error típ.	Significación	Intervalo de confianza al 95%. Límite inferior	Límite superior
DHS de Tukey	Administrativo	Seguridad	-3100,35	2011,232	,272	-7829,14	1628,44
		Directivo	-36139,26*	1220,747	,000	-39009,47	-33269,05
	Seguridad	Administrativo	3100,35	2011,232	,272	-1628,44	7829,14
		Directivo	-33038,91*	2230,514	,000	-38283,28	-27794,54
	Directivo	Administrativo	36139,26*	1220,747	,000	33269,05	39009,47
		Seguridad	33038,91*	2230,514	,000	27794,54	38283,28
Scheffe	Administrativo	Seguridad	-3100,35	2011,232	,306	-8039,13	1838,44
		Directivo	-36139,26*	1220,747	,000	-39136,93	-33141,59
	Seguridad	Administrativo	3100,35	2011,232	,306	-1838,44	8039,13
		Directivo	-33038,91*	2230,514	,000	-38516,16	-27561,65
	Directivo	Administrativo	36139,26*	1220,747	,000	33141,59	39136,93
		Seguridad	33038,91*	2230,514	,000	27561,65	38516,16
Bonferroni	Administrativo	Seguridad	-3100,35	2011,232	,372	-7932,57	1731,87
		Directivo	-36139,26*	1220,747	,000	-39072,25	-33206,27
	Seguridad	Administrativo	3100,35	2011,232	,372	-1731,87	7932,57
		Directivo	-33038,91*	2230,514	,000	-38397,98	-27679,84
	Directivo	Administrativo	36139,26*	1220,747	,000	33206,27	39072,25
		Seguridad	33038,91*	2230,514	,000	27679,84	38397,98
Tamhane	Administrativo	Seguridad	-3100,35*	568,679	,000	-4483,07	-1717,63
		Directivo	-36139,26*	2029,912	,000	-41078,30	-31200,21
	Seguridad	Administrativo	3100,35*	568,679	,000	1717,63	4483,07
		Directivo	-33038,91*	2031,840	,000	-37982,78	-28095,04
	Directivo	Administrativo	36139,26*	2029,912	,000	31200,21	41078,30
		Seguridad	33038,91*	2031,840	,000	28095,04	37982,78
T3 de Dunnet	Administrativo	Seguridad	-3100,35*	568,679	,000	-4482,16	-1718,53

Figura 8-80

Los diagramas de dispersión por nivel (Figuras 8-81 y 8-82) corroboran la heterogeneidad de varianzas (sus puntos no están alineados). El gráfico de los residuos (Figura 8-83) muestra que, aunque los residuos parecen independientes (no muestran una pauta de variación sistemática) la dispersión de los mismos no es la misma a lo largo de todos los niveles pronosticados (varianzas no homogéneas). La gráfica de valores observados contra pronosticados muestra tendencia lineal, luego no hay problemas de linealidad en el modelo. La Figura 8-84 presenta el gráfico de perfil de categoría laboral por clasificación étnica, cuyos puntos son medias de salario halladas en el subgrupo resultante de combinar cada nivel de *catlab* con cada nivel de *minoría*.

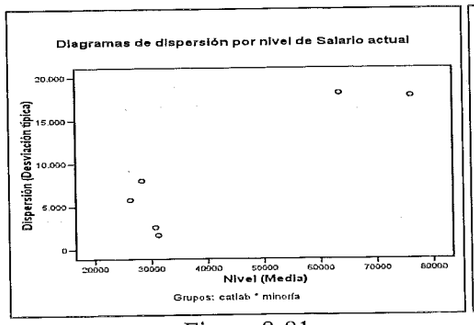


Figura 8-81

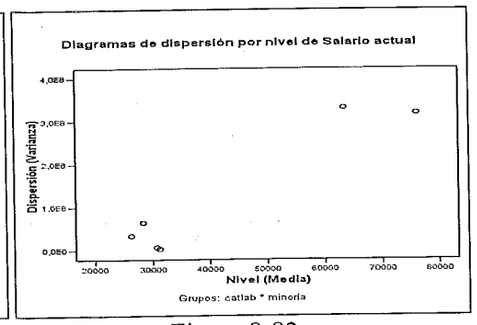


Figura 8-82

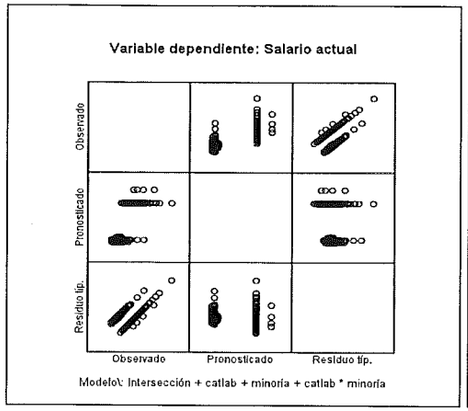


Figura 8-83

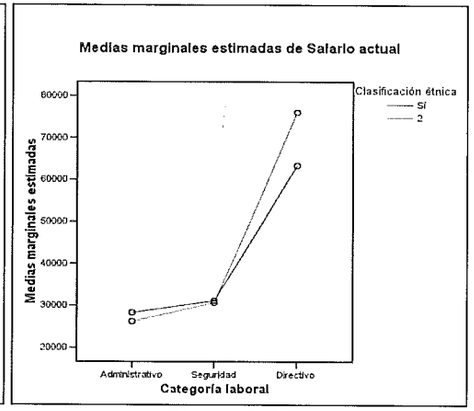


Figura 8-84

Para realizar la segunda parte del ejercicio ampliamos el modelo con las covariables experiencia previa (*expprev*) y tiempo desde el contrato (*tiempemp*) rellenando la pantalla de entrada del procedimiento *MLG Univariante* como se indica en la Figura 8-85. Se rellena el botón *Opciones* según la Figura 8-86 y al pulsar *Aceptar* se obtiene la tabla ANCOVA de la Figura 8-87. Como las covariables tienen p-valor menor que 0,05, se encuentran linealmente relacionadas con la variable salario. También se observa que la significación de minoría ha aumentado hasta el 92%, por lo que la introducción de las covariables mejora el modelo. La Figura 8-88 muestra las estimaciones del modelo con covariables.

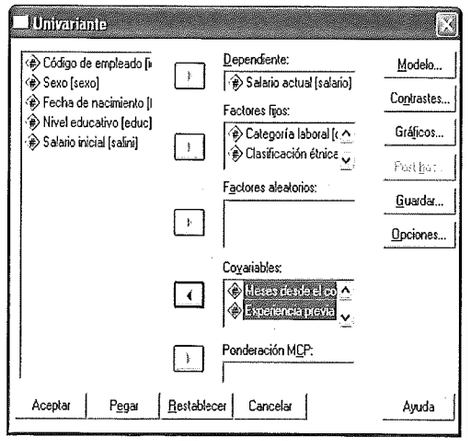


Figura 8-85

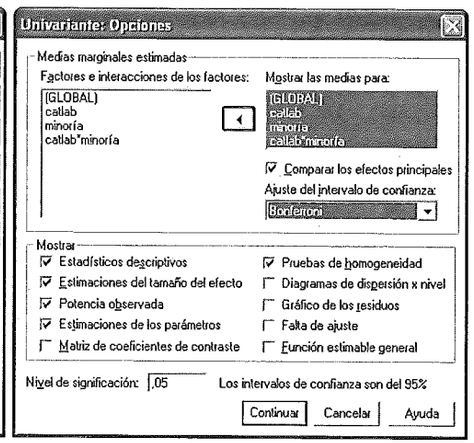


Figura 8-86

Contraste de Levene sobre la igualdad de las varianzas error[a]

Variable dependiente: Salario actual

F	gl1	gl2	Significación
24,469	5	468	,000

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intercept+tiempemp+expprev+catlab+minoría+catlab * minoría

Pruebas de los efectos inter-sujetos

Variable dependiente: Salario actual

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación	Eta al cuadrado parcial	Parámetro de no centralidad	Potencia observada[a]
Modelo corregido	9,205E+10[b]	7	1,31E+10	133,588	,000	,667	935,116	1,000
Intersección	6558230593	1	6,56E+09	66,626	,000	,125	66,626	1,000
tiempemp	1210038518	1	1,21E+09	12,293	,000	,026	12,293	,938
expprev	492908058	1	4,93E+08	5,008	,026	,011	5,008	,608
catlab	2,720E+10	2	1,36E+10	138,181	,000	,372	276,361	1,000
minoría	299670835	1	3,00E+08	3,044	,082	,006	3,044	,414
catlab * minoría	1091353985	2	5,46E+08	5,544	,004	,023	11,087	,853
Error	4,587E+10	466	98433306					
Total	6,995E+11	474						
Total corregida	1,379E+11	473						

a. Calculado con alfa = ,05

b. R cuadrado = ,667 (R cuadrado corregida = ,662)

Figura 8-87

Estimaciones de los parámetros

Variable dependiente: Salario actual

Parámetro	B	Error típ.	t	Significación	Intervalo de confianza al 95%: Límite inferior	Límite superior	Eta al cuadrado parcial	Parámetro de no centralidad	Potencia observada[a]
Intersección	66527,537	5930,510	11,218	,000	54873,683	78181,391	,213	11,218	1,000
tiempemp	160,464	45,767	3,506	,000	70,529	250,399	,026	3,506	,938
expprev	-11,323	5,060	-2,238	,026	-21,266	-1,380	,011	2,238	,608
[catlab=1]	-52333,1	5113,150	-10,235	,000	-62380,812	-42285,4	,184	10,235	1,000
[catlab=2]	-46338,5	5725,869	-8,093	,000	-57590,191	-35086,8	,123	8,093	1,000
[catlab=3]	0[b]	.	.	.	.	.	.	.	.
[minoría=1]	-15428,0	5126,476	-3,009	,003	-25501,833	-5354,088	,019	3,009	,852
[minoría=2]	0[b]	.	.	.	.	.	.	.	.
[catlab=1] * [minoría=1]	17530,738	5273,216	3,324	,001	7168,512	27892,964	,023	3,324	,913
[catlab=1] * [minoría=2]	0[b]	.	.	.	.	.	.	.	.
[catlab=2] * [minoría=1]	17430,840	6455,398	2,700	,007	4745,545	30116,135	,015	2,700	,769
[catlab=2] * [minoría=2]	0[b]	.	.	.	.	.	.	.	.
[catlab=3] * [minoría=1]	0[b]	.	.	.	.	.	.	.	.
[catlab=3] * [minoría=2]	0[b]	.	.	.	.	.	.	.	.

a. Calculado con alfa = ,05

b. Al parámetro se le ha asignado el valor cero porque es redundante.

Figura 8-88

Ejercicio 8-2. En 26 muestras de cerámica con distintos orígenes se mide el porcentaje de óxido de 5 metales (Aluminio, Hierro, Magnesio, Calcio y Sodio) según los siguientes datos:

Origen	Al	Fe	Mg	Ca	Na	Origen	Al	Fe	Mg	Ca	Na
Llanederyn	14.4	7.00	4.30	0.15	0.51	Llanederyn	13.8	7.08	3.43	0.12	0.17
Llanederyn	14.6	7.09	3.88	0.13	0.20	Llanederyn	11.5	6.37	5.64	0.16	0.14
Llanederyn	13.8	7.06	5.34	0.20	0.20	Llanederyn	10.9	6.26	3.47	0.17	0.22
Llanederyn	10.1	4.26	4.26	0.20	0.18	Llanederyn	11.6	5.78	5.91	0.18	0.16
Llanederyn	11.1	5.49	4.52	0.29	0.30	Llanederyn	13.4	6.92	7.23	0.28	0.20
Llanederyn	12.4	6.13	5.69	0.22	0.54	Llanederyn	13.1	6.64	5.51	0.31	0.24
Llanederyn	12.7	6.69	4.45	0.20	0.22	Llanederyn	12.5	6.44	3.94	0.22	0.23
Caldicot	11.8	5.44	3.94	0.30	0.04	Caldicot	11.6	5.39	3.77	0.29	0.06
IslandThorns	18.3	1.28	0.67	0.03	0.03	IslandThorns	15.8	2.39	0.63	0.01	0.04
IslandThorns	18.0	1.50	0.67	0.01	0.06	IslandThorns	18.0	1.88	0.68	0.01	0.04
IslandThorns	20.8	1.51	0.72	0.07	0.10	AshleyRails	17.7	1.12	0.56	0.06	0.06
AshleyRails	18.3	1.14	0.67	0.06	0.05	AshleyRails	16.7	0.92	0.53	0.01	0.05
AshleyRails	14.8	2.74	0.67	0.03	0.05	AshleyRails	19.1	1.64	0.60	0.10	0.03

Realizar un análisis multivariante de la varianza para medir las diferencias entre las características químicas de cerámica proveniente de distintos orígenes en el Reino Unido. Adicionalmente, es conveniente conocer si la cerámica proveniente de Llanederyn difiere de las restantes.

Se trata de un análisis multivariante de la varianza con un factor y cinco variables dependientes. Para llevarlo a cabo introducimos los datos en el fichero 8-2.*sav* y elegimos *Analizar* → *Modelo lineal general* → *Multivariante* (Figura 8-89) rellenando la pantalla de entrada como se indica en la Figura 8-90 (factor *origen* y variables dependientes *Al, Fe, Mg, Ca* y *Na.* Rellenamos el botón *Post hoc* como se indica en la Figura 8-91 y el botón *Opciones* según la Figura 8-92. Al pulsar *Aceptar* se obtiene la salida.

La Figura 8-93 muestra diversos contrastes multivariados sobre la pertinencia del ajuste cuyos p-valores bajos en general admiten el ajuste. También se muestra el estadístico de Levenne para la igualdad de varianzas del error (según su p-valor, se acepta salvo para el caso del magnesio). La Figura 8-94 muestra la tabla resumen del MANOVA, conteniendo fuentes de variación, sumas de cuadrados, grados de libertad, medias cuadráticas, estadísticos F y sus p-valores, referido todo ello a los efectos presentes en nuestro modelo de un factor. También presenta medidas de la calidad del ajuste como la eta-cuadrado y otras medidas como el parámetro de no centralidad y la potencia de los contrastes. En general la significación de los parámetros del modelo resulta bastante alta (p-valores pequeños, coeficientes eta-cuadrado altos y potencias altas). Las estimaciones de los parámetros del modelo multivariante de la varianza presentan en general p-valores aceptables (Figura 8-95).

Las comparaciones por pares entre las medias de los grupos delimitados en las variables dependientes por el factor *origen* (Figura 8-96) muestran que para el valor *Llanederyn* no hay significatividad de las diferencias de medias con el valor *Caldicot*, pero sí con *IslandThorns* y *AshleyRails*. Lo mismo se deduce de los contrastes *post hoc* de comparaciones múltiples (Figura 8-97). Por tanto la cerámica proveniente de Llanederyn difiere de la proveniente de *IslandThorns* y *AshleyRails*, pero no de la proveniente de *Caldicot*.

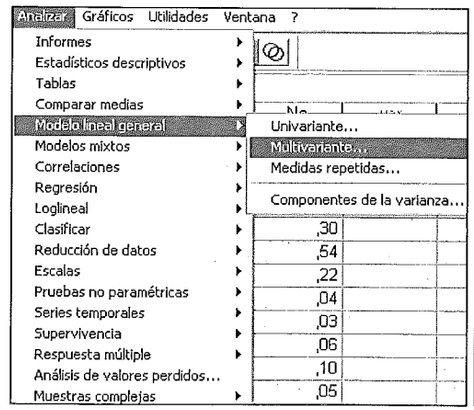


Figura 8-89

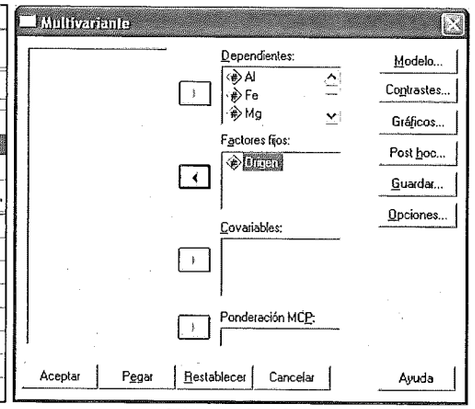


Figura 8-90

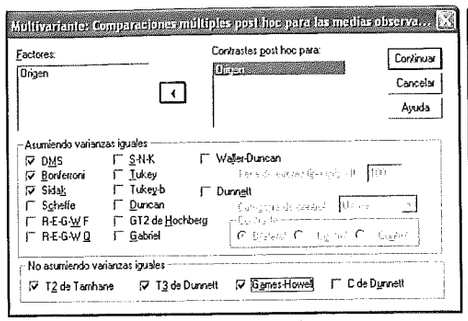


Figura 8-91

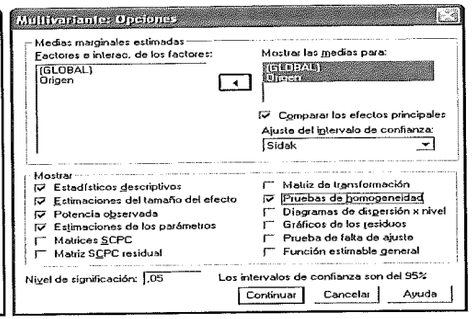


Figura 8-92

Contrastes multivariados[d]

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Significación	Eta al cuadrado parcial	Parámetro de no centralidad	Potencia observada[a]
Intersección	Traza de Pillai	,989	319,930[b]	5,000	18,000	,000	,989	1599,650	1,000
	Lambda de Wilks	,011	319,930[b]	5,000	18,000	,000	,989	1599,650	1,000
	Traza de Hotelling	88,869	319,930[b]	5,000	18,000	,000	,989	1599,650	1,000
	Raíz mayor de Roy	88,869	319,930[b]	5,000	18,000	,000	,989	1599,650	1,000
Origen	Traza de Pillai	1,554	4,298	15,000	60,000	,000	,518	64,476	1,000
	Lambda de Wilks	,012	13,089	15,000	50,091	,000	,769	166,904	1,000
	Traza de Hotelling	35,439	39,376	15,000	50,000	,000	,922	590,646	1,000
	Raíz mayor de Roy	34,161	136,644[c]	5,000	20,000	,000	,972	683,222	1,000

a. Calculado con alfa = ,05
b. Estadístico exacto
c. El estadístico es un límite superior para la F el cual ofrece un límite inferior para el nivel de significación.
d. Diseño: Intercept+Origen

Contraste de Levene sobre la igualdad de las varianzas error[a]

	F	gl1	gl2	Significación
Al	,916	3	22	,449
Fe	1,035	3	22	,396
Mg	9,242	3	22	,000
Ca	1,755	3	22	,185
Na	2,641	3	22	,075

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.
a. Diseño: Intercept+Origen

Figura 8-93

Pruebas de los efectos inter-sujetos

Fuente	Variable dependiente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación	Eta al cuadrado parcial	Parámetro de no centralidad	Potencia observada[a]
Modelo corregido	Al	175,610[b]	3	58,537	26,669	,000	,784	80,008	1,000
	Fe	134,222[c]	3	44,741	89,883	,000	,925	269,648	1,000
	Mg	103,351[d]	3	34,450	49,120	,000	,870	147,360	1,000
	Ca	,205[e]	3	,068	29,157	,000	,799	87,470	1,000
	Na	,258[f]	3	,086	9,503	,000	,564	28,508	,990
Intersección	Al	3676,822	1	3676,822	1675,154	,000	,987	1675,154	1,000
	Fe	231,962	1	231,962	466,006	,000	,955	466,006	1,000
	Mg	102,149	1	102,149	145,647	,000	,869	145,647	1,000
	Ca	,341	1	,341	145,505	,000	,869	145,505	1,000
	Na	,167	1	,167	18,430	,000	,456	18,430	,984
Origen	Al	175,610	3	58,537	26,669	,000	,784	80,008	1,000
	Fe	134,222	3	44,741	89,883	,000	,925	269,648	1,000
	Mg	103,351	3	34,450	49,120	,000	,870	147,360	1,000
	Ca	,205	3	,068	29,157	,000	,799	87,470	1,000
	Na	,258	3	,086	9,503	,000	,564	28,508	,990
Error	Al	48,288	22	2,195					
	Fe	10,951	22	,498					
	Mg	15,430	22	,701					
	Ca	,051	22	,002					
	Na	,199	22	,009					
Total	Al	5684,600	26						
	Fe	664,140	26						
	Mg	375,381	26						
	Ca	,814	26						
	Na	1,110	26						
Total corregida	Al	223,898	25						
	Fe	145,172	25						
	Mg	118,780	25						

Figura 8-94

Estimaciones de los parámetros

Variable dependiente	Parámetro	B	Error típ.	t	Significación	Intervalo de confianza al 95% Límite inferior	Límite superior	Eta al cuadrado parcial	Parámetro de no centralidad	Potencia observada[a]
Al	Intersección	17,320	,663	26,141	,000	15,946	18,694	,969	26,141	1,000
	[Origen=1]	-4,756	,772	-6,161	,000	-6,356	-3,155	,633	6,161	1,000
	[Origen=2]	-5,620	1,240	-4,534	,000	-8,191	-3,049	,483	4,534	,991
	[Origen=3]	,860	,937	,918	,369	-1,083	2,803	,037	,918	,142
	[Origen=4]	0[b]	.	.	.	.	.	.	.	.
Fe	Intersección	1,612	,316	4,792	,000	,958	2,166	,511	4,792	,996
	[Origen=1]	4,860	,369	13,222	,000	4,099	5,622	,888	13,222	1,000
	[Origen=2]	3,903	,590	6,612	,000	2,679	5,127	,665	6,612	1,000
	[Origen=3]	,200	,446	,448	,658	-,725	1,125	,009	,448	,071
	[Origen=4]	0[b]	.	.	.	.	.	.	.	.
Mg	Intersección	,606	,375	1,618	,120	-,171	1,383	,106	1,618	,340
	[Origen=1]	4,220	,436	9,673	,000	3,316	5,125	,810	9,673	1,000
	[Origen=2]	3,249	,701	4,637	,000	1,796	4,702	,494	4,637	,993
	[Origen=3]	,068	,530	,128	,899	-1,030	1,166	,001	,128	,052
	[Origen=4]	0[b]	.	.	.	.	.	.	.	.
Ca	Intersección	,052	,022	2,404	,025	,007	,097	,208	2,404	,632
	[Origen=1]	,150	,025	5,957	,000	,098	,202	,617	5,957	1,000
	[Origen=2]	,243	,040	6,004	,000	,159	,327	,621	6,004	1,000
	[Origen=3]	-,026	,031	-,850	,405	-,089	,037	,032	,850	,129
	[Origen=4]	0[b]	.	.	.	.	.	.	.	.
Na	Intersección	,048	,043	1,128	,272	-,040	,136	,055	1,128	,190
	[Origen=1]	,203	,050	4,088	,000	,100	,306	,432	4,088	,974
	[Origen=2]	,002	,080	,025	,980	-,163	,167	,000	,025	,050
	[Origen=3]	,006	,060	,100	,922	-,119	,131	,000	,100	,051
	[Origen=4]	0[b]	.	.	.	.	.	.	.	.

Figura 8-95

Comparaciones por pares

Variable dependiente	(I) Origen	(J) Origen	Diferencia entre medias (I-J)	Error típ.	Significación[a]	Intervalo de confianza al 95 % para diferencia[a] Límite inferior	Límite superior
Al	Lanederyn	Caldicot	,864	1,120	,972	-2,371	4,100
		IslandThorns	-5,616*	,772	,000	-7,846	-3,386
		AshleyRalls	-4,756*	,772	,000	-6,986	-2,526
	Caldicot	Lanederyn	-,864	1,120	,972	-4,100	2,371
		IslandThorns	-6,480*	1,240	,000	-10,061	-2,899
		AshleyRalls	-5,620*	1,240	,001	-9,201	-2,039
	IslandThorns	Lanederyn	5,616*	,772	,000	3,386	7,846
		Caldicot	6,480*	1,240	,000	2,899	10,061
		AshleyRalls	,860	,937	,937	-1,847	3,567
	AshleyRalls	Lanederyn	4,756*	,772	,000	2,526	6,986
		Caldicot	5,620*	1,240	,001	2,039	9,201
		IslandThorns	-,860	,937	,937	-3,567	1,847
Fe	Lanederyn	Caldicot	,957	,533	,419	-,584	2,498
		IslandThorns	4,860*	,368	,000	3,598	5,722
		AshleyRalls	4,860*	,368	,000	3,798	5,922
	Caldicot	Lanederyn	-,957	,533	,419	-2,498	,584
		IslandThorns	3,703*	,590	,000	1,998	5,408
		AshleyRalls	3,903*	,590	,000	2,198	5,608
	IslandThorns	Lanederyn	-4,860*	,368	,000	-5,722	-3,598
		Caldicot	-3,703*	,590	,000	-5,408	-1,998
		AshleyRalls	,200	,446	,998	-1,089	1,489
	AshleyRalls	Lanederyn	-4,860*	,368	,000	-5,922	-3,798
		Caldicot	-3,903*	,590	,000	-5,608	-2,198
		IslandThorns	-,200	,446	,998	-1,489	1,089
Mg	Lanederyn	Caldicot	,971	,633	,593	-,858	2,801
		IslandThorns	4,152*	,436	,000	2,892	5,413

Figura 8-96

Comparaciones múltiples

Bonferroni

Variable dependiente	(I) Origen	(J) Origen	Diferencia entre medias (I-J)	Error típ.	Significación	Intervalo de confianza al 95%. Límite inferior	Límite superior
Al	Lanederyn	Caldicot	,8643	1,11993	1,000	-2,3819	4,1104
		IslandThorns	-5,6157*	,77186	,000	-7,8530	-3,3785
		AshleyRalls	-4,7557*	,77186	,000	-6,9930	-2,5185
	Caldicot	Lanederyn	-,8643	1,11993	1,000	-4,1104	2,3819
		IslandThorns	-6,4800*	1,23953	,000	-10,0728	-2,8872
		AshleyRalls	-5,6200*	1,23953	,001	-9,2128	-2,0272
	IslandThorns	Lanederyn	5,6157*	,77186	,000	3,3785	7,8530
		Caldicot	6,4800*	1,23953	,000	2,8872	10,0728
		AshleyRalls	,8600	,93700	1,000	-1,8559	3,5759
	AshleyRalls	Lanederyn	4,7557*	,77186	,000	2,5185	6,9930
		Caldicot	5,6200*	1,23953	,001	2,0272	9,2128
		IslandThorns	-,8600	,93700	1,000	-3,5759	1,8559
Fe	Lanederyn	Caldicot	,9571	,53333	,519	-,5887	2,5030
		IslandThorns	4,6601*	,36757	,000	3,5947	5,7256
		AshleyRalls	4,8601*	,36757	,000	3,7947	5,9256
	Caldicot	Lanederyn	-,9571	,53333	,519	-2,5030	,5887
		IslandThorns	3,7030*	,59028	,000	1,9920	5,4140
		AshleyRalls	3,9030*	,59028	,000	2,1920	5,6140
	IslandThorns	Lanederyn	-4,6601*	,36757	,000	-5,7256	-3,5947
		Caldicot	-3,7030*	,59028	,000	-5,4140	-1,9920
		AshleyRalls	,2000	,44621	1,000	-1,0934	1,4934
	AshleyRalls	Lanederyn	-4,8601*	,36757	,000	-5,9256	-3,7947
		Caldicot	-3,9030*	,59028	,000	-5,6140	-2,1920
		IslandThorns	-,2000	,44621	1,000	-1,4934	1,0934
Mg	Lanederyn	Caldicot	,9714	,63306	,835	-,8635	2,8064
		IslandThorns	4,1524*	,43631	,000	2,8976	5,4171

Figura 8-97

Resolvemos ahora el mismo ejercicio con el software SAS. La sintaxis puede ser la siguiente:

```
data ceramica;
      title1 "Ceramica inglesa";
      input Origen $ Al Fe Mg Ca Na @@;
      datalines;
Llanederyn    14.4 7.00 4.30 0.15 0.51   Llanederyn    13.8 7.08 3.43 0.12 0.17
Llanederyn    14.6 7.09 3.88 0.13 0.20   Llanederyn    11.5 6.37 5.64 0.16 0.14
Llanederyn    13.8 7.06 5.34 0.20 0.20   Llanederyn    10.9 6.26 3.47 0.17 0.22
Llanederyn    10.1 4.26 4.26 0.20 0.18   Llanederyn    11.6 5.78 5.91 0.18 0.16
Llanederyn    11.1 5.49 4.52 0.29 0.30   Llanederyn    13.4 6.92 7.23 0.28 0.20
Llanederyn    12.4 6.13 5.69 0.22 0.54   Llanederyn    13.1 6.64 5.51 0.31 0.24
Llanederyn    12.7 6.69 4.45 0.20 0.22   Llanederyn    12.5 6.44 3.94 0.22 0.23
Caldicot      11.8 5.44 3.94 0.30 0.04   Caldicot      11.6 5.39 3.77 0.29 0.06
IslandThorns  18.3 1.28 0.67 0.03 0.03   IslandThorns  15.8 2.39 0.63 0.01 0.04
IslandThorns  18.0 1.50 0.67 0.01 0.06   IslandThorns  18.0 1.88 0.68 0.01 0.04
IslandThorns  20.8 1.51 0.72 0.07 0.10   AshleyRails   17.7 1.12 0.56 0.06 0.06
AshleyRails   18.3 1.14 0.67 0.06 0.05   AshleyRails   16.7 0.92 0.53 0.01 0.05
AshleyRails   14.8 2.74 0.67 0.03 0.05   AshleyRails   19.1 1.64 0.60 0.10 0.03
;
   proc glm data=ceramica;
      class origen;
      model Al Fe Mg Ca Na = origen;
      contrast 'Llanederyn contra el resto' origen 1 1 1 -3;
      manova h=_all_ / printe printh;
   run;
```

La salida es la siguiente:

```
                          Ceramica inglesa
                        The GLM Procedure

                       Class Level Information

              Class        Levels    Values

              Origen            4    AshleyRa Caldicot IslandTh Llaneder

                     Number of observations    26

Dependent Variable: Al

                                        Sum of
      Source                    DF      Squares     Mean Square    F Value    Pr > F

      Model                      3   175.6103187      58.5367729      26.67    <.0001
      Error                     22    48.2881429       2.1949156
      Corrected Total           25   223.8984615

                R-Square     Coeff Var      Root MSE      Al Mean
                0.784330      10.22284      1.481525     14.49231

      Source                    DF     Type I SS     Mean Square    F Value    Pr > F
      Origen                     3   175.6103187      58.5367729      26.67    <.0001

      Source                    DF    Type III SS    Mean Square    F Value    Pr > F
      Origen                     3   175.6103187      58.5367729      26.67    <.0001

   Contrast                     DF    Contrast SS    Mean Square    F Value    Pr > F
   Llanederyn contra el resto    1   58.58336640    58.58336640      26.69    <.0001
```



```
Dependent Variable: Fe

                                        Sum of
      Source                    DF      Squares     Mean Square    F Value    Pr > F

      Model                      3   134.2216158      44.7405386      89.88    <.0001
      Error                     22    10.9508457       0.4977657
      Corrected Total           25   145.1724615

                R-Square     Coeff Var      Root MSE       Fe Mean
                0.924567      15.79171      0.705525      4.467692

      Source                    DF     Type I SS     Mean Square    F Value    Pr > F
      Origen                     3   134.2216158      44.7405386      89.88    <.0001

      Source                    DF    Type III SS    Mean Square    F Value    Pr > F
      Origen                     3   134.2216158      44.7405386      89.88    <.0001

   Contrast                     DF    Contrast SS    Mean Square    F Value    Pr > F
   Llanederyn contra el resto    1   71.15144132    71.15144132     142.94    <.0001

Dependent Variable: Mg

                                        Sum of
      Source                    DF      Squares     Mean Square    F Value    Pr > F

      Model                      3   103.3505270      34.4501757      49.12    <.0001
      Error                     22    15.4296114       0.7013460
      Corrected Total           25   118.7801385

                R-Square     Coeff Var      Root MSE       Mg Mean
                0.870099      26.65777      0.837464      3.141538

      Source                    DF     Type I SS     Mean Square    F Value    Pr > F
      Origen                     3   103.3505270      34.4501757      49.12    <.0001

      Source                    DF    Type III SS    Mean Square    F Value    Pr > F
      Origen                     3   103.3505270      34.4501757      49.12    <.0001

   Contrast                     DF    Contrast SS    Mean Square    F Value    Pr > F
   Llanederyn contra el resto    1   56.59349339    56.59349339      80.69    <.0001

Dependent Variable: Ca

                                        Sum of
      Source                    DF      Squares     Mean Square    F Value    Pr > F

      Model                      3    0.20470275      0.06823425      29.16    <.0001
      Error                     22    0.05148571      0.00234026
      Corrected Total           25    0.25618846

                R-Square     Coeff Var      Root MSE       Ca Mean
                0.799032      33.01265      0.048376      0.146538

      Source                    DF     Type I SS     Mean Square    F Value    Pr > F
      Origen                     3    0.20470275      0.06823425      29.16    <.0001

      Source                    DF    Type III SS    Mean Square    F Value    Pr > F
      Origen                     3    0.20470275      0.06823425      29.16    <.0001

   Contrast                     DF    Contrast SS    Mean Square    F Value    Pr > F
   Llanederyn contra el resto    1    0.03531688     0.03531688      15.09    0.0008

Dependent Variable: Na

                                        Sum of
      Source                    DF      Squares     Mean Square    F Value    Pr > F

      Model                      3    0.25824560      0.08608187       9.50    0.0003
      Error                     22    0.19929286      0.00905877
```

Corrected Total	25	0.45753846	

R-Square	Coeff Var	Root MSE	Na Mean
0.564424	60.06350	0.095178	0.158462

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Origen	3	0.25824560	0.08608187	9.50	0.0003

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Origen	3	0.25824560	0.08608187	9.50	0.0003

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Llanederyn contra el resto	1	0.23344446	0.23344446	25.77	<.0001

The GLM Procedure
Multivariate Analysis of Variance
E = Error SSCP Matrix

	Al	Fe	Mg	Ca	Na
Al	48.288142857	7.0800714286	0.6080142857	0.1064714286	0.5889571429
Fe	7.0800714286	10.950845714	0.5270571429	-0.155194286	0.0667585714
Mg	0.6080142857	0.5270571429	15.429611429	0.4353771429	0.0276157143
Ca	0.1064714286	-0.155194286	0.4353771429	0.0514857143	0.0100785714
Na	0.5889571429	0.0667585714	0.0276157143	0.0100785714	0.1992928571

Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|

DF = 22	Al	Fe	Mg	Ca	Na
Al	1.000000	0.307889	0.022275	0.067526	0.189853
		0.1529	0.9196	0.7595	0.3856
Fe	0.307889	1.000000	0.040547	-0.206685	0.045189
	0.1529		0.8543	0.3440	0.8378
Mg	0.022275	0.040547	1.000000	0.488478	0.015748
	0.9196	0.8543		0.0180	0.9431
Ca	0.067526	-0.206685	0.488478	1.000000	0.099497
	0.7595	0.3440	0.0180		0.6515
Na	0.189853	0.045189	0.015748	0.099497	1.000000
	0.3856	0.8378	0.9431	0.6515	

H = Type III SSCP Matrix for Origen

	Al	Fe	Mg	Ca	Na
Al	175.61031868	-149.295533	-130.8097066	-5.889163736	-5.372264835
Fe	-149.295533	134.22161582	117.74503516	4.8217865934	5.3259491209
Mg	-130.8097066	117.74503516	103.35052703	4.2091613187	4.7105458242
Ca	-5.889163736	4.8217865934	4.2091613187	0.2047027473	0.154782967
Na	-5.372264835	5.3259491209	4.7105458242	0.154782967	0.2582456044

Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for Origen
E = Error SSCP Matrix

Characteristic Root	Percent	Characteristic Vector V'EV=1 Al	Fe	Mg	Ca	Na
34.1611140	96.39	0.09562211	-0.26330469	-0.05305978	-1.87982100	-0.47071123
1.2500994	3.53	0.02651891	-0.01239715	0.17564390	-4.25929785	1.23727668
0.0275396	0.08	0.09082220	0.13159869	0.03508901	-0.15701602	-1.39364544
0.0000000	0.00	0.03673984	-0.15129712	0.20455529	0.54624873	-0.17402107
0.0000000	0.00	0.06862324	0.03056912	-0.10662399	2.51151978	1.23668841

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall Origen Effect
H = Type III SSCP Matrix for Origen
E = Error SSCP Matrix

S=3 M=0.5 N=8

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.01230091	13.09	15	50.091	<.0001
Pillai's Trace	1.55393619	4.30	15	60	<.0001
Hotelling-Lawley Trace	35.43875302	40.59	15	29.13	<.0001
Roy's Greatest Root	34.16111399	136.64	5	20	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

H = Contrast SSCP Matrix for Llanederyn contra el resto

	Al	Fe	Mg	Ca	Na
Al	58.583366402	-64.56230291	-57.57983466	-1.438395503	-3.698102513
Fe	-64.56230291	71.151441323	63.456352116	1.5851961376	4.0755256878
Mg	-57.57983466	63.456352116	56.593493386	1.4137558201	3.6347541005
Ca	-1.438395503	1.5851961376	1.4137558201	0.0353168783	0.0907993915
Na	-3.698102513	4.0755256878	3.6347541005	0.0907993915	0.2334444577

Characteristic Roots and Vectors of: E Inverse * H, where
H = Contrast SSCP Matrix for Llanederyn contra el resto
E = Error SSCP Matrix

Characteristic Root	Percent	Characteristic Vector V'EV=1 Al	Fe	Mg	Ca	Na
16.1251646	100.00	-0.08883488	0.25458141	0.08723574	0.98158668	0.71925759
0.0000000	0.00	-0.00503538	0.03825743	-0.17632854	5.16256699	-0.01022754
0.0000000	0.00	0.00162771	-0.08885364	-0.01774069	-0.83096817	2.17644566
0.0000000	0.00	0.04450136	-0.15722494	0.22156791	0.00000000	0.00000000
0.0000000	0.00	0.11939206	0.10833549	0.00000000	0.00000000	0.00000000

MANOVA Test Criteria and Exact F Statistics for the Hypothesis
of No Overall Llanederyn contra el resto Effect
H = Contrast SSCP Matrix for Llanederyn contra el resto
E = Error SSCP Matrix

S=1 M=1.5 N=8

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.05839360	58.05	5	18	<.0001
Pillai's Trace	0.94160640	58.05	5	18	<.0001
Hotelling-Lawley Trace	16.12516462	58.05	5	18	<.0001
Roy's Greatest Root	16.12516462	58.05	5	18	<.0001

La opción PRINTE ha originado que se imprima la matriz de sumas de cuadrados del error y productos cruzados, cuyos elementos de la diagonal son las sumas de los cuadrados del error para los correspondientes análisis univariantes. Esta opción también imprime la matriz de correlaciones asociada con la anterior. Según los resultados, los elementos más correlacionados son el óxido de magnesio y el óxido de calcio (r=0,488). La opción PRINTH imprime la matriz SSCP para contrastar hipótesis. Los elementos de la diagonal de esta matriz son las sumas de cuadrados del modelo para los análisis univariantes.

Se computan 4 contrastes multivariantes basados en raíces y vectores característicos de la matriz $E^{-1}H$. Los cuatro tests dan el mismo resultado (diferencias significativas entre las composiciones químicas), ya que sólo existe un grado de libertad. Según los tests existe una diferencia global entre la composición química de las muestras provenientes de diferentes orígenes, lo que constata la diferencia derivada de cada contraste univariante. Además las muestras provenientes de Llanederyn son distintas de la media de las muestras restantes. Obsérvese que todos los p-valores son menores que una milésima, tanto para constrastes univariantes como multivariantes.

Ejercicio 8-3. Se trata de estudiar si la edad y el género están relacionados con el sentido y el olfato. Para ello, un total de 180 sujetos de 20 a 89 años son expuestos a 40 diferentes olores, y para cada olor se pregunta a los sujetos cuál de entre cuatro palabras describe mejor el olor. Los datos se dividen en 5 grupos de edades (0,25], (25,40], (40,55], (55,70] y (70, más) y a cada grupo se le asigna como índice olfativo la proporción de olores correctamente identificados. Los datos son los siguientes:

```
1 1.381  1 1.322  1 1.162  1 1.275  1 1.381  1 1.275  1 1.322
1 1.492  1 1.322  1 1.381  1 1.162  1 1.013  1 1.322  1 1.322
1 1.275  1 1.492  1 1.322  1 1.322  1 1.492  1 1.322  1 1.381
1 1.234  1 1.162  1 1.381  1 1.381  1 1.381  1 1.322  1 1.381
1 1.322  1 1.381  1 1.275  1 1.492  1 1.275  1 1.322  1 1.275
1 1.381  1 1.234  1 1.105
2 1.234  2 1.234  2 1.381  2 1.322  2 1.492  2 1.234  2 1.381
2 1.381  2 1.492  2 1.492  2 1.275  2 1.492  2 1.381  2 1.492
2 1.322  2 1.275  2 1.275  2 1.275  2 1.322  2 1.492  2 1.381
2 1.322  2 1.492  2 1.196  2 1.322  2 1.275  2 1.234  2 1.322
2 1.098  2 1.322  2 1.381  2 1.275  2 1.492  2 1.492  2 1.381
2 1.196
3 1.381  3 1.381  3 1.492  3 1.492  3 1.492  3 1.098  3 1.492
3 1.381  3 1.234  3 1.234  3 1.129  3 1.069  3 1.234  3 1.322
3 1.275  3 1.230  3 1.234  3 1.234  3 1.322  3 1.322  3 1.381
4 1.322  4 1.381  4 1.381  4 1.322  4 1.234  4 1.234  4 1.234
4 1.381  4 1.322  4 1.275  4 1.275  4 1.492  4 1.234  4 1.098
4 1.322  4 1.129  4 0.687  4 1.322  4 1.322  4 1.234  4 1.129
4 1.492  4 0.810  4 1.234  4 1.381  4 1.040  4 1.381  4 1.381
4 1.129  4 1.492  4 1.129  4 1.098  4 1.275  4 1.322  4 1.234
4 1.196  4 1.234  4 0.585  4 0.785  4 1.275  4 1.322  4 0.712
4 0.810
5 1.322  5 1.234  5 1.381  5 1.275  5 1.275  5 1.322  5 1.162
5 0.909  5 0.502  5 1.234  5 1.322  5 1.196  5 0.859  5 1.196
5 1.381  5 1.322  5 1.234  5 1.275  5 1.162  5 1.162  5 0.585
5 1.013  5 0.960  5 0.662  5 1.129  5 0.531  5 1.162  5 0.737
5 1.098  5 1.162  5 1.040  5 0.558  5 0.960  5 1.098  5 0.884
5 1.162  5 1.098  5 0.859  5 1.275  5 1.162  5 0.785  5 0.859
```

Realizar un contraste de igualdad de índices olfativos para los distintos grupos de edad y de la variabilidad de estos índices por grupos de edad.

Se trata de realizar un contraste de igualdad de medias y de varianzas de índice olfativo por grupos de edad en un modelo univariante de la varianza utilizando, por ejemplo, los contrastes Levenne (igualdad de varianzas) y Welch (igualdad de medias). La sintaxis será la siguiente:

```
data datos;
     input grupoedad indiceolfativo @@;
     datalines;
1 1.381  1 1.322  1 1.162  1 1.275  1 1.381  1 1.275  1 1.322
1 1.492  1 1.322  1 1.381  1 1.162  1 1.013  1 1.322  1 1.322
1 1.275  1 1.492  1 1.322  1 1.322  1 1.492  1 1.322  1 1.381
1 1.234  1 1.162  1 1.381  1 1.381  1 1.381  1 1.322  1 1.381
1 1.322  1 1.381  1 1.275  1 1.492  1 1.275  1 1.322  1 1.275
1 1.381  1 1.234  1 1.105
2 1.234  2 1.234  2 1.381  2 1.322  2 1.492  2 1.234  2 1.381
2 1.381  2 1.492  2 1.492  2 1.275  2 1.492  2 1.381  2 1.492
2 1.322  2 1.275  2 1.275  2 1.275  2 1.322  2 1.492  2 1.381
2 1.322  2 1.492  2 1.196  2 1.322  2 1.275  2 1.234  2 1.322
2 1.098  2 1.322  2 1.381  2 1.275  2 1.492  2 1.492  2 1.381
2 1.196
3 1.381  3 1.381  3 1.492  3 1.492  3 1.492  3 1.098  3 1.492
3 1.381  3 1.234  3 1.234  3 1.129  3 1.069  3 1.234  3 1.322
3 1.275  3 1.230  3 1.234  3 1.234  3 1.322  3 1.322  3 1.381
4 1.322  4 1.381  4 1.381  4 1.322  4 1.234  4 1.234  4 1.234
```



```
4 1.381  4 1.322  4 1.275  4 1.275  4 1.492  4 1.234  4 1.098
4 1.322  4 1.129  4 0.687  4 1.322  4 1.322  4 1.234  4 1.129
4 1.492  4 0.810  4 1.234  4 1.381  4 1.040  4 1.381  4 1.381
4 1.129  4 1.492  4 1.129  4 1.098  4 1.275  4 1.322  4 1.234
4 1.196  4 1.234  4 0.585  4 0.785  4 1.275  4 1.322  4 0.712
4 0.810
5 1.322  5 1.234  5 1.381  5 1.275  5 1.275  5 1.322  5 1.162
5 0.909  5 0.502  5 1.234  5 1.322  5 1.196  5 0.859  5 1.196
5 1.381  5 1.322  5 1.234  5 1.275  5 1.162  5 1.162  5 0.585
5 1.013  5 0.960  5 0.662  5 1.129  5 0.531  5 1.162  5 0.737
5 1.098  5 1.162  5 1.040  5 0.558  5 0.960  5 1.098  5 0.884
5 1.162  5 1.098  5 0.859  5 1.275  5 1.162  5 0.785  5 0.859
;
proc glm data=datos;
   class grupoedad;
   model indiceolfativo = grupoedad;
   means grupoedad / hovtest welch;
run;
```

La salida es la siguiente:

The GLM Procedure

Class Level Information

Class	Levels	Values
grupoedad	5	1 2 3 4 5

Number of observations 180
Dependent Variable: indiceolfativo

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2.13878141	0.53469535	16.65	<.0001
Error	175	5.61970399	0.03211259		
Corrected Total	179	7.75848539			

R-Square	Coeff Var	Root MSE	indiceolfativo Mean
0.275670	14.52664	0.179200	1.233594

Source	DF	Type I SS	Mean Square	F Value	Pr > F
grupoedad	4	2.13878141	0.53469535	16.65	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grupoedad	4	2.13878141	0.53469535	16.65	<.0001

Levene's Test for Homogeneity of indiceolfativo Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
grupoedad	4	0.0799	0.0200	6.35	<.0001
Error	175	0.5503	0.00314		

Welch's ANOVA for indiceolfativo

Source	DF	F Value	Pr > F
grupoedad	4.0000	13.72	<.0001
Error	78.7489		

Level of grupoedad	N	--------indiceolfativo------- Mean	Std Dev
1	38	1.31689474	0.10365373
2	36	1.34513889	0.10574232
3	21	1.30614286	0.12748933
4	43	1.20109302	0.21961425
5	42	1.05961905	0.24594635

Los p-valores de los contrastes de la F, tanto para la igualdad de medias como para la igualdad de varianzas son todos menores que una milésima, lo que indica diferencias significativas al 99% entre medias de índice olfativo por grupos de edad (*Welch's ANOVA*) y entre las variabilidades de estas medias (*Levene's Test for Homogeneity*), es decir, entre las varianzas de índices olfativos por grupos de edad.

Ejercicio 8-4. Consideramos el fichero panelocde.sav que contiene variables de 10 países de la OCDE. Se trata de ajustar con estos datos un modelo de panel que explique la esperanza de vida (Ev) en función del producto nacional bruto (PNB), el gasto social (GastoS) y la protección social (Cobertura) utilizando un modelo de panel con coeficientes constantes. Realizar los ajustes con efectos fijos y aleatorios.

En primer ligar realizaremos el ajuste con efectos fijos. Se trata de ajustar mediante un modelo mixto el modelo de panel de efectos fijos siguiente:

$$Ev = Constante + a*PNB + b*GastoS + c*Cobertura + Efectos\ fijos + Error$$

Elegimos *Analizar* → *Modelos mixtos* → *Lineal* (Figura 8-98) y en la primera pantalla (Figura 8-99) hacemos clic en *Continuar*. A continuación rellenamos la pantalla de entrada del procedimiento como se indica en la Figura 8-100. Los botones *Efectos fijos*, *Estadísticos* y *Medias* se rellenan según las Figura 8-101 a 8-103. Al hacer clic en *Continuar* y *Aceptar* se obtiene el ajuste de las Figuras 8-104 y 8-105. Se observa que este ajuste es equivalente al obtenido en el problema anterior (significatividades y estimaciones de parámetros y efectos fijos) ya que el test de Levenne verificó variabilidad constante del error en los diferentes países. El modelo de panel con efectos fijos ajustado es:

$$Ev = 65{,}234 + 0{,}323\ PNB - 0{,}207\ GastoS + 0{,}076\ Cobertura + Efectos\ fijos + e$$

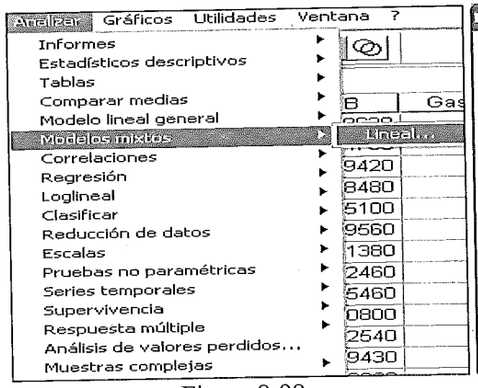


Figura 8-98

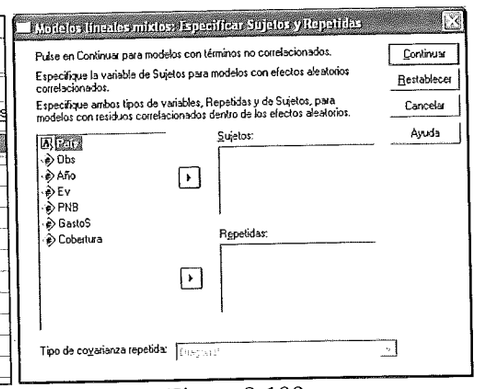


Figura 8-199

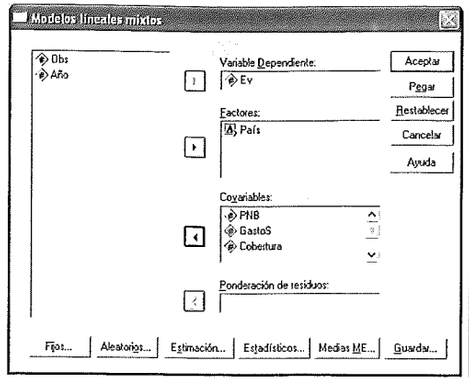


Figura 8-100

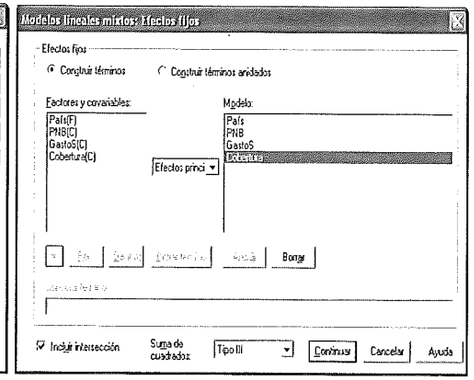


Figura 8-101

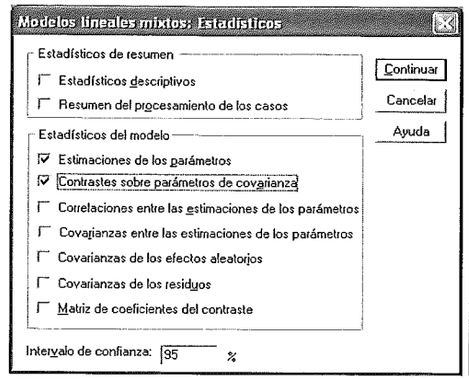


Figura 8-102

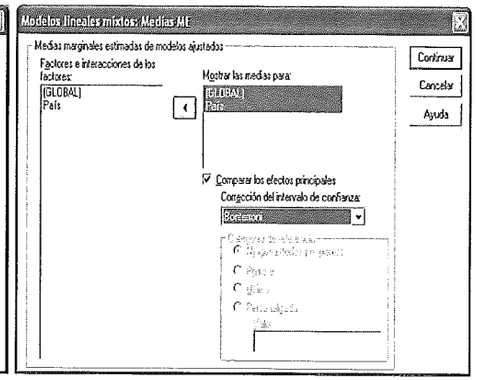


Figura 8-103

Análisis de modelos mixtos

Dimensión de modelo[a]

		Número de niveles	Número de parámetros
Efectos fijos	Intersección	1	1
	País	10	9
	PNB	1	1
	GastoS	1	1
	Cobertura	1	1
Residuos			1
Total		14	14

a. Variable dependiente: Ev.

Criterios de información[a]

-2 log de la verosimilitud restringida	101,906
Criterio de información de Akaike (AIC)	103,906
Criterio de Hurvich y Tsai (AICC)	104,066
Criterio de Bozdogan (CAIC)	106,202
Criterio bayesiano de Schwarz (BIC)	105,202

Los criterios de información se muestran en formatos de mejor cuanto más pequeños.

a. Variable dependiente: Ev.

Figura 8-104

Tipo III de contrastes de efectos fijos[a]

Origen	Numerador df	Denominador df	Valor F	Sig.
Intersección	1	27	2240,627	,000
País	9	27	3,683	,004
PNB	1	27	13,630	,001
GastoS	1	27	,019	,892
Cobertura	1	27	20,698	,000

a. Variable dependiente: Ev.

Estimaciones de efectos fijos[b]

Parámetro	Estimación	Error típico	gl	t	Sig.	Intervalo de confianza 95% Límite inferior	Límite superior
Intersección	65,2343846	1,4727652	27	44,293	,000	62,2124790	68,25629
[País=Australia]	-1,3168868	,7066936	27	-1,863	,073	-2,7669023	,1331287
[País=Austria]	-1,9318398	,6859593	27	-2,816	,009	-3,3393119	-,5243673
[País=Bélgica]	-,9498286	,6855303	27	-1,385	,177	-2,3552207	,4569634
[País=Canadá]	-,1222560	,6869229	27	-,178	,860	-1,5317054	1,2871934
[País=Dinamarca]	-,2528007	,7035121	27	-,359	,722	-1,6952863	1,1906870
[País=EE.UU.]	2,8740575	1,1080763	27	2,594	,015	,6004727	5,1476422
[País=España]	1,0286333	,7143382	27	1,441	,161	-,4360677	2,4953342
[País=Reino Un.]	-1,3575799	,7113488	27	-1,908	,067	-2,8171470	,1019872
[País=Suecia]	,4462355	,7244442	27	,616	,543	-1,0402012	1,9326722
[País=Suiza]	0[a]	0	.	.	.	.	.
PNB	,3234695	,0869804	27	3,719	,001	,1450005	,5019384
GastoS	-,2071627	1,5153191	27	-,137	,892	-3,3163407	2,9020154
Cobertura	,0756400	,0166259	27	4,550	,000	,0415266	,1097535

a. Se ha establecido este parámetro en cero porque es redundante.

b. Variable dependiente: Ev.

Figura 8-105

Para ajustar el modelo con efectos aleatorios elegimos *Analizar* → *Modelos mixtos* → *Lineal* y en la primera pantalla hacemos clic en *Continuar*. A continuación rellenamos la pantalla de entrada del procedimiento como en el caso de efectos fijos. Los botones *Efectos fijos*, *Aleatorios*, *Estadísticos* y *Estimación* se rellenan según las Figuras 8-106 a 8-109. Al hacer clic en *Continuar* y *Aceptar* se obtiene el ajuste de las Figuras 8-1110 y 8-111. Se observa que este ajuste difiere del obtenido en el problema anterior (significatividades y estimaciones de parámetros). El modelo de panel con efectos aleatorios ajustado es:

$$Ev = 67,447 + 0,383\ PNB - 0,628\ GastoS + 0,045\ Cobertura + e$$

Se observa que las diferencias en las estimaciones de los parámetros son muy pequeñas respecto del modelo de efectos fijos. La Figura 8-112 muestra también las estimaciones de las componentes de la varianza.

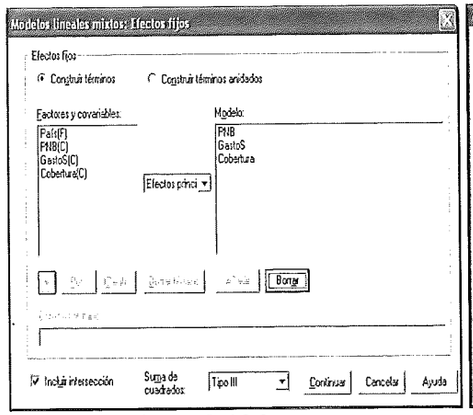


Figura 8-106

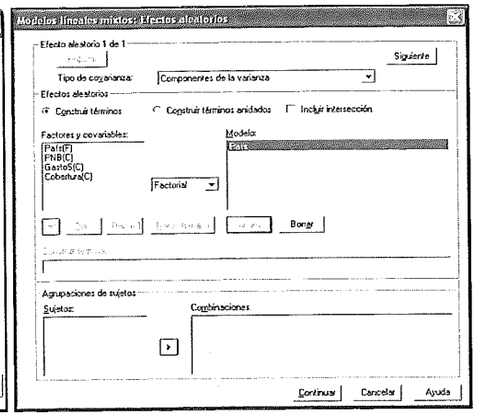


Figura 8-107

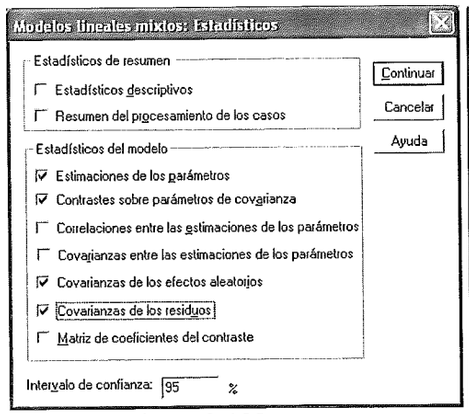


Figura 8-108

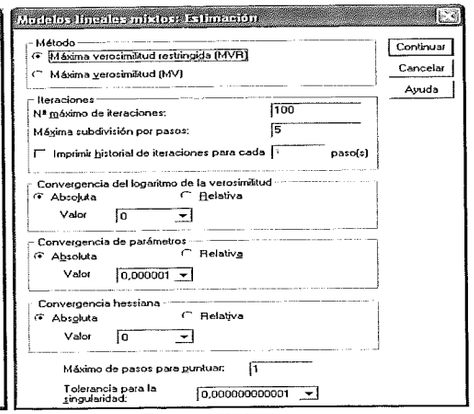


Figura 8-109

Dimensión de modelo[b]

		Número de niveles	Estructura de covarianza	Número de parámetros
Efectos fijos	Intersección	1		1
	PNB	1		1
	GastoS	1		1
	Cobertura	1		1
Efectos aleatorios	País[a]	10	Componentes de la varianza	1
Residuos				1
Total		14		6

a. Como en la versión 11.5, han cambiado las reglas de sintaxis para el subcomando RANDOM. Su sintaxis de comandos puede generar resultados que difieran de los generados por versiones anteriores. Si utiliza la sintaxis de SPSS 11, consulte el manual de referencia de la sintaxis actual para obtener más información al respecto.

b. Variable dependiente: Ev.

Criterios de información[a]

-2 log de la verosimilitud restringida	132,957
Criterio de información de Akaike (AIC)	136,957
Criterio de Hurvich y Tsai (AICC)	137,320
Criterio de Bozdogan (CAIC)	142,124
Criterio bayesiano de Schwarz (BIC)	140,124

Los criterios de información se muestran en formatos de mejor cuanto más pequeños.

a. Variable dependiente: Ev.

Figura 8-110

Tipo III de contrastes de efectos fijos[a]

Origen	Numerador df	Denominador df	Valor F	Sig.
Intersección	1	20,310	3997,418	,000
PNB	1	27,740	18,878	,000
GastoS	1	27,092	,164	,689
Cobertura	1	20,436	14,210	,001

a. Variable dependiente: Ev.

Estimaciones de efectos fijos[a]

Parámetro	Estimación	Error típico	gl	t	Sig.	Intervalo de confianza 95% Límite inferior	Límite superior
Intersección	67,4478233	1,0667881	20,310	63,225	,000	65,2247148	69,67093
PNB	,3832705	,0882112	27,740	4,345	,000	,2025018	,5640392
GastoS	-,6286228	1,5525824	27,092	-,405	,689	-3,8137514	2,5565058
Cobertura	,0457460	,0121355	20,436	3,770	,001	,0204654	,0710256

a. Variable dependiente: Ev.

Parámetros de covarianza

Estimaciones de parámetros de covarianza[a]

Parámetro		Estimación	Error típico	Wald Z	Sig.	Intervalo de confianza 95% Límite inferior	Límite superior
Residuos		1,0105924	,2955106	3,420	,001	,5697357	1,7925805
País	Varianza	,7103003	,6137652	1,157	,247	,1305945	3,8633048

a. Variable dependiente: Ev.

Figura 8-111